Covid-19: Time Series Analysis

[1]Er. Gaurisha Gupta, [2]Dr. Rahul Malhotra, [3]Er. Kamal Kumar

[1]Student, [2]Director, [3]H.O.D
Department of Computer Science,
Swami Devi Dyal Institute of Engineering and Technology, Kurukshetra University, Haryana, India

***Abstract*: Since December end 2019, an outbreak of a novel coronavirus disease (COVID-19; previously known as 2019-nCoV) was reported in Wuhan, China, which has subsequently affected 210 countries worldwide. In general, Coronavirus is an acute resolved disease, but it can be deadly also, with a 6.5% case fatality rate. Various diseases onset might cause death due to massive damage and progressive respiratory failure. As of 26 September 2020, data from the World Health Organization (WHO) have shown that more than 5.9 Million confirmed cases have been identified in 210 countries/regions. On 30 January 2020, the World Health Organization declared that COVID-19 as the sixth public health emergency of international concern. In such a circumstance, Artificial Intelligence and machine learning can assume an immense job in foreseeing a flare-up and limiting or slowing down its spread.**

In this thesis, our objective is how Artificial Intelligence and Machine Learning can play a enormous role in predicting an outbreak and also minimizing or impede its spread. With machine learning,

We used the ARIMA model to the time series data of confirmed COVID-19 cases in India. Autocorrelation function (ACF) graph and partial autocorrelation (PACF) graph is used to find the initial parameters of ARIMA models. These ARIMA models are then tested for variance in normality and stationery through the collection of data. With this model, we try to learn more from the past than from what we think the mechanism [of transmission] is. With mechanism approaches, people try to build models that are based on an understanding of how epidemics spread. Machine learning (Time Series) algorithms to track it, and quickly realized that lending the use of our technology to the global public is the minimum we can do to help during this very difficult time.

In this thesis, We also focus on Artificial Intelligence which is good at combing through the collection of data to find connections that make it easier to determine which kinds of treatments could work or which experiments to pursue next. In addition, this thesis will also define how AI also can help in healthcare technologies to detect and fight against coronavirus.

***Keywords*: Autocorrelation function (ACF), Partial autocorrelation (PACF), Autoregressive Integrated Moving-average method.**

I. INTRODUCTION

According to the World Health Organization (WHO), viral diseases appear and represent a serious issue to public health. In the last twenty-five years, several viral pandemics like SARS-CoV in 2002 to 2003, and H1N1 influenza in 2009, have been recorded, and recently, In 2012the Middle East respiratory syndrome coronavirus (MERS-CoV) was first detected in Saudi Arabia. And Now the novel coronavirus disease (COVID-19) was reported in Wuhan, China, which has subsequently affected 210 countries worldwide. The novel coronavirus disease (COVID-19) has created enormous chaos around the world, affecting most people's lives and causing a large number of deaths. Its first cases were tested positively in Wuhan, China in December 2019 and now it has been spread to almost every country. Governments of many different countries have proposed policies to mitigate the impacts of the COVID-19 pandemic. Peoples who infected with coronavirus may have little or no symptoms. You may not know the symptoms because they resemble the common cold or flu. The symptoms of people with COVID-19 ranged from mild symptoms to serious illnesses. Symptoms like fever, cough, chills, repeated shaking with chills, difficulty breathing, headache, sore throat, muscle aches, new taste or smell, bluish lips or face, new confusion or inability to be awakened, and persistent chest pain or pressure.

COVID-19 can generally spread through direct or indirect contact, droplet sprays for short-range transmission and aerosol for long-distance transmission. Coronavirus disease 2019 is believed to support the belief that the virus spreads between people in close contact with each other through respiratory droplets that arise when an infected person coughs, sneezes, or speaks, possibly from people can be inhaled nearby and through contact with contaminated surfaces or objects. Coronavirus can also spread by people who haven't any symptoms.

After breathing out-produce these droplets, they generally fall onto the ground or the surfaces rather than remain in the air over long distances. People may also be infected by touching an infected surface and then touching their eyes, nose, or mouth. The virus can spread and live on any surface for more than 72 hours. During the first three days, It is most contagious after the onset of symptoms, although spread may be possible before symptoms appear and in later stages of the disease. In general, Coronavirus is an acute disease, but it can also be deadly, with a 6.5% case fatality rate. As of 26 September 2020, data from the World Health Organization (WHO) have shown that more than 5 million cases have been identified confirmed in 210 countries. About 97,000 people have died from COVID-19, while More than 5.2 million have been confirmed in at least 188 countries and territories. More than 2.1 million people have recovered to date. Every Asian country and European countries are affected by a coronavirus, while more European countries are witnessing a surge in cases. India replaced China as the country with the highest coronavirus cases, while more than 92% of the global coronavirus cases are currently outside China. Fig 1: represents the graph of ten most affected countries with coronavirus.

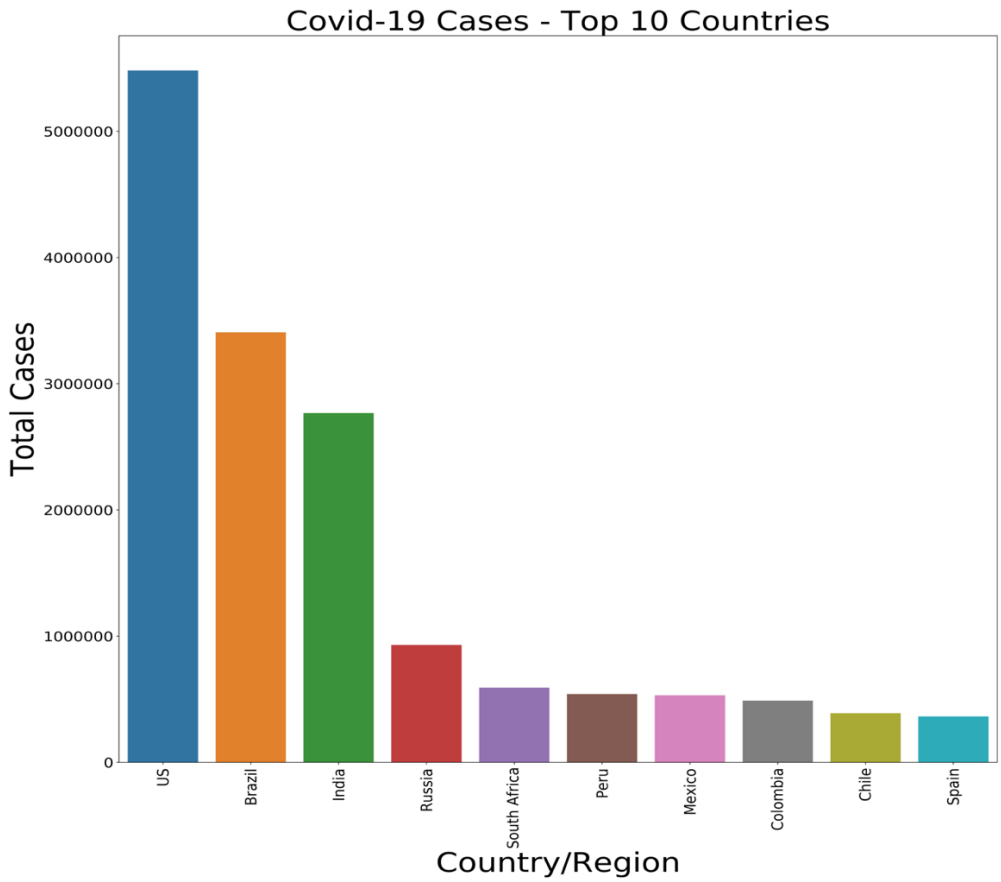

Figure 1: Top Ten most affected countries with coronavirus

On January 30 2020, the World Health Organization (WHO) declared COVID-19 as the sixth public health emergency of international concern. The large number of asymptomatic cases and the possibility of people remaining positive for the virus even after a recovery, are of international concern, as many countries seek to finish the lockdowns and start economic activity as the spread of the virus declines. Coronavirus is continuing to spread globally, and death toll has surpassed 48,000. COVID-19 confirmed cases and deaths continue to increase rapidly throughout Italy and the U.S. The World Health Organization (WHO) said Wednesday, April 1, that in the last five weeks saw "rapidly growth in the number of new COVID-19 cases, reaching and spreading in almost every country, territory and area. Presently, the highest confirmed cases of coronavirus infections have been reported in US, however, the cases are abruptly rising in Spain, Italy, France and Germany daily. China, the place where the diseases arise, is now receiving a very few cases.4 The first case of coronavirus infection in India was reported on 30 January 2020 in Kerala, which was an imported case from Wuhan city of China. In the initial phase the spread was slow and only 3 people were positive for more than a month. However, the number started rising exponentially after one month and continue to do so. The numbers in India have reached up to 10,453 for confirmed COVID-19 infected cases with 358 deaths and 1181 recoveries as reported on 13 April 2020.At present, there isn't any treatment and a vaccination for the COVID-19 infection. Currently, it is a major health crisis in all over the world and we can also say that it is 'an enemy to humanity'. In this circumstance, the only chance is preventing the occurrence of infection and preparing our healthcare system for the up-comings problems. In that reference, it is very crucial to create models that are computationally competent as well as realistic so that they can help in policy makers, medical issues and also general public. Providing future forecast and modeling the disease of possible number of daily cases can assist the medical system in getting prepared for the new patients. The statistical prediction models can be used in forecasting as well as controlling the global epidemic threat. In the present effort, we have applied Auto-Regressive Integrated Moving Average (ARIMA) model for predicting the incidence of 2019-nCov disease. As compared to other prediction models, for instance support vector machine (SVM) and wavelet neural network (WNN), ARIMA model is more capable and batter in the prediction of natural adversities. 10 For our study, we have identified the best ARIMA model and then predicted the number the cases for the next 20 days. The main objective of the study is to choose the best predictive model and apply it to forecast future incidence of COVID-19 cases in India.

II. Rational of the Study

The coronavirus disease (COVID-19) has created tremendous chaos around the world and affecting many people's lives also causing a large number of deaths. Its first cases were detected in Wuhan, China in December 2019 and now it has been spread to almost every country. Governments of many different countries around the world have declared policies to mitigate the impacts of the COVID-19 pandemic. Science and technology have contributed significantly to the implementation of these policies during this unprecedented and chaotic time. For example, various robots are used in hospitals to deliver food, medicine, and also used for

treatments of coronavirus patients or drones are used to disinfect streets or public spaces. Many researchers, scientists and doctors are rushing to produce drugs and medicines to treat infected patients while many doctors are attempting to investigate vaccines to prevent the virus.

The Objective of this is how to fight with Covid-19 with AI/ML. Artificial Intelligence (AI) and Machine Learning is a potentially powerful tool in the fight against the COVID-19 pandemic. The main purpose of this thesis is to predict the future infected cases to support prevention of the disease and aid in the healthcare service preparation. Following that notion, we have created a model and then predict it for forecasting future COVID-19 cases in India. The study indicates an increasing trend for the covid cases in the coming days. A time series analysis also presents an increasing trend in the number of cases. It is supposed that the present prediction models will assist the government and medical personnel to be prepared for the upcoming conditions and have more readiness in healthcare systems. AI can, for present purposes, be defined as Machine Learning (ML), Natural Language Processing (NLP), and Computer Vision applications to teach computers to use big data-based models for pattern recognition, explanation, and prediction. These functions can be useful to, predict, explain (treat) and recognize (diagnose) COVID-19 infections, and help manage socio-economic impacts.

III. METHODOLOGY

DATASET

Confirmed cases of COVID-19 infection are collected for India as well as countries with highest confirmed infection (US, Brajil, India, France, Russia, China and Iran) and countries in South-East Asia region (India, Indonesia, Thailand, Bangladesh, Sri Lanka, Maldives, Nepal, Bhutan and Timor-Leste), as per World Health Organization region classification, from the official website of Johns Hopkins University. (https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html) from 22 January 2020 to 18 August 2020. This data is used to build predictive models.

	Province/State	Country/Region	Lat	Long	Total Cases	Date	Cases
0		Afghanistan	33.93911	67.709953	37599	1/22/20	0
1		Albania	41.1533	20.1683	7654	1/22/20	0
2		Algeria	28.0339	1.6596	39444	1/22/20	0
3		Andorra	42.5063	1.5218	1005	1/22/20	0
4		Angola	-11.2027	17.8739	1966	1/22/20	0
5		Antigua and Barbu	17.0608	-61.7964	93	1/22/20	0
6		Argentina	-38.4161	-63.6167	305966	1/22/20	0
7		Armenia	40.0691	45.0382	41846	1/22/20	0
8	Australian Capital Ter	Australia	-35.4735	149.0124	113	1/22/20	0
9	New South Wales	Australia	-33.8688	151.2093	3966	1/22/20	0
10	Northern Territory	Australia	-12.4634	130.8456	33	1/22/20	0
11	Queensland	Australia	-27.4698	153.0251	1092	1/22/20	0
12	South Australia	Australia	-34.9285	138.6007	462	1/22/20	0
13	Tasmania	Australia	-42.8821	147.3272	230	1/22/20	0
14	Victoria	Australia	-37.8136	144.9631	17446	1/22/20	0
15	Western Australia	Australia	-31.9505	115.8605	647	1/22/20	0
16		Austria	47.5162	14.5501	23829	1/22/20	0
17		Azerbaijan	40.1431	47.5769	34474	1/22/20	0
18		Bahamas	25.025885	-78.035889	1424	1/22/20	0
19		Bahrain	26.0275	50.55	47581	1/22/20	0

Figure 2. Covid-19 Dataset

MODEL DEVELOPMENT

The ARIMA is a class of models which also known as the Box-Jenkins method. The Box- Jenkins method related to the fitting of a mixed ARIMA model to a given data set. ARIMA, short form for 'Auto-Regressive Integrated Moving Average' is a class of models which explain a given time series based situations on its past values, i.e., its own lags and the lagged forecast errors, so that this equation can be used for predict future values. An ARIMA model is characterized by 3 terms: a, b, c.

Where,

a is the order of the AR term.

b is the order of the MA term.

c is the number of differencing required to make the time series stationary.

ARIMA model:

- Recognition of the model: This includes Recognize the most suitable values for the components of the AR and MA, and deciding whether the variable needs first differentiation to induce stationary. The Auto Correlation function (ACF) and the Partial Auto Correlation function (PACF) are used to determine the best model.
- Estimation: This usually involves the use of a least-squares estimation process.

- Diagnostic testing: This usually is the test for autocorrelation. If this part fails, then the process returns to the identification section and begins again, usually by the addition of extra variables.
- Forecasting: The ARIMA models are useful for forecasting due to the use of past variables.

We have applied an ARIMA model to the time series data of confirmed COVID-19 cases in India. Autocorrelation function graph and partial autocorrelation graph is used to find the lags of ARIMA models. These ARIMA models are then used for variance in stationary and normality. Next, they are checked for accuracy by observing their MAPE, MAD and MSD values to determine the finest model to forecast. In addition, the best fit ARIMA model is compared with Linear Trend, Quadratic Trend, S-Curve Trend, Moving Average, Single Exponential as well as Double Exponential models using an output of measure of accuracy, viz. MAPE, MAD, MSD, so as to select the finest model to forecast. The finest model is the one which has the lowest value for all the measures. After fitting the model, its parameters are estimated followed by verification of the model. This model is employed to forecast confirmed COVID-19. The model for forecasting future confirmed COVID-19 cases is represented as,

$$ARIMA(a,b,c): Xt=\alpha 1Xt-1+\alpha 2Xt-2+\beta 1Zt-1+\beta 2Zt-2+Zt \quad (1)$$

where $Zt=Xt-Xt-1$

Here, Xt is the predicted number of confirmed Coronavirus confirmed cases at tth day, α1, α2, β1 and β2 are parameters whereas Zt is the residual term for tth day. The trend of forthcoming incidences can be estimated from the previous cases and a time series analysis is performed for this purpose. Time series forecasting refers to the employment of a model to forecast future data based on previously observed data.11 In the present study, time series analysis is used to recognize the trends in confirmed COVID-19 cases in India over the period of 22 January 2020 to 13 April 2020 and to predict future cases from 14 April 2020 till 3 May 2020. The level of statistical significance of the built model is set at 0.05.

A comparative study is also performed to examine the status of confirmed COVID-19 cases of India with respect to those of highly infected countries. A similar comparison is made with the countries of South-East Asia region as well. All the model developments, computations and comparisons have been performed using Minitab software.

Autoregressive models are used when the current level of the series is thought to depend on the recent history of the series. An autoregressive model of order a (AR(a)) or ARIMA (a, 0, 0) where is the autoregressive parameter. In practice Autoregressive models higher than order 2 are rarely observed.

The following Fig. 3 show the typical ACF and PACF for stationary AR(1) and AR(2).

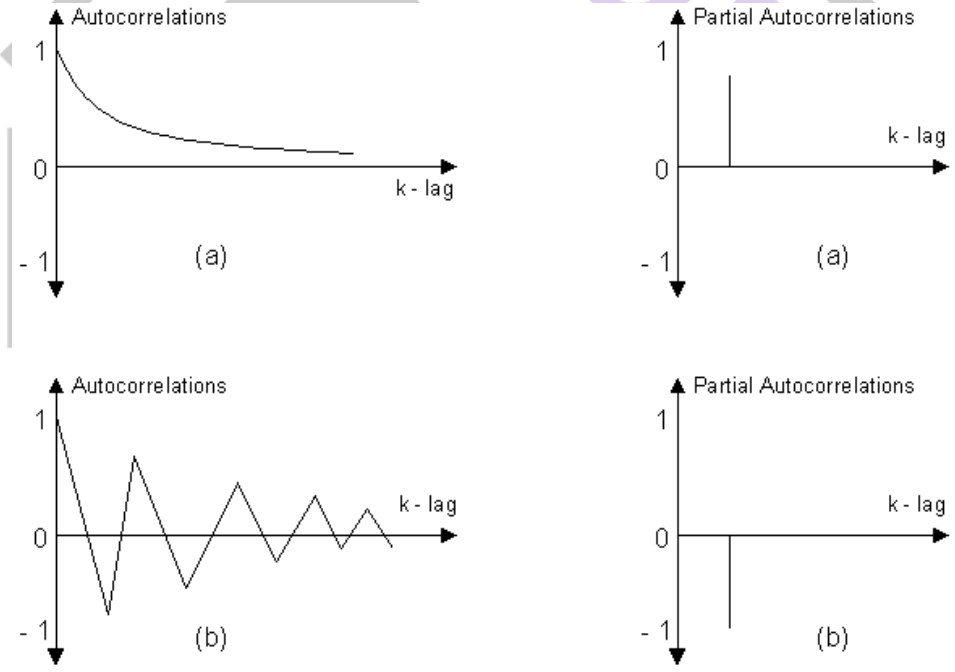

Figure 3. Typical autocorrelation and partial autocorrelation functions for stationary

The most common approach to make a series stationary is to subtract the last value from the present value. Sometimes more than one differentiation required, depending on the complexity of the series. Therefore, the value of c is the minimum amount of differentiation needed to render the sequence stationary, and if the time series is stationary already, then c = 0. The 'a' is the order of 'Auto-Regressive' (AR) term; it refers to the number of Y lags which should be used as predictors. The 'b' is the order of the (MA) 'Moving Average' term; it refers to the number of lagged errors in the forecast that should go into the ARIMA model. The objective of the fitting ARIMA model is to perfectly recognize the stochastic mechanism of the time series and predict or forecast future values.

Such types of approaches have also proven useful in other types of various scenarios in which models are used for discrete-time series and dynamic systems are created. This method is, however, not appropriate for seasonal series or lead times with a broad random variable. This time series model has been used in the current study to predict or forecast the number of COVID-19 confirmed cases in India. The steps of the ARIMA model building methodology are presented in a flowchart below in Fig. 4.

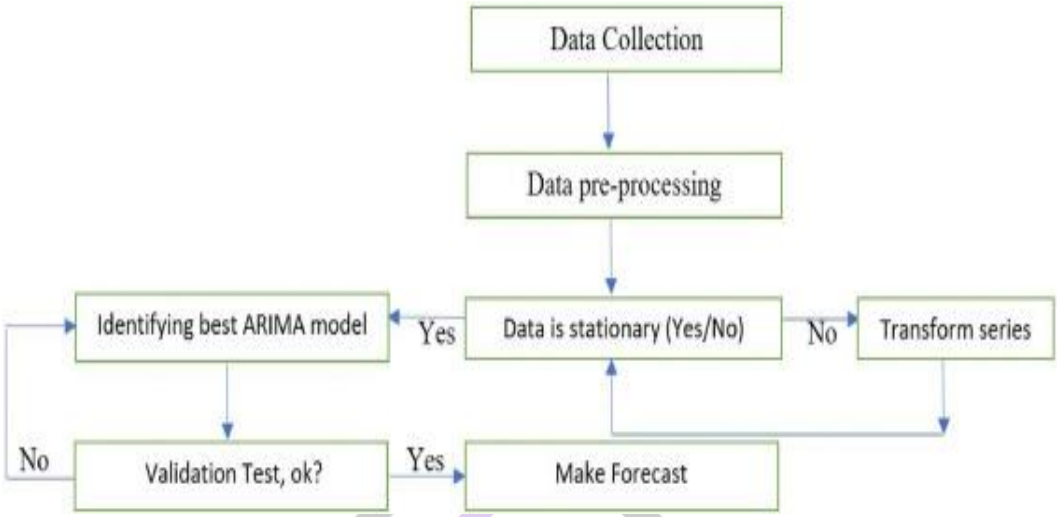

Figure 4. Methodology to apply the ARIMA model for forecasting.

Once the model will identified, and the model parameter can be estimated, then the model is determined with a different set of parameters and start building the models. It is basically identify with the assumption that the model about the random error is satisfied using statistical diagnostic tests and residual plots that can be used to analyze the suitability of various models to historical data. The mode can be selected on the basis of values of specific criteria like Normalized Bayesian Information Criteria (BIC).

IV. Time Series Analysis

Introduction to Time Series

The term "time series" itself, define a data storing format, which consists of the two mandatory components - time units and the corresponding value assigned for the given time unit. Values of the time series need to denote the same meaning and correlate among the nearby values. The restriction is that at the same time there can be at most single value for every time unit. For example, sequences, which just enumerate some values, they do not fulfil the time series requirements. In theory, there are two types fundamental ways, how time series data are recorded. The first way, values are measured just for the specific timestamps, what may occur occasionally, or periodically according to concrete conditions, but anyway, result will be a discrete set of values, formally called discrete time series. This is a very common case and frequently observed in practice. In economy sector, most of the indicators are measured periodically with the specific time periods, therefore economic indicators represent an appropriate example of discrete time series. The second option is, that data are measured and recorded continuously along with the time intervals. Electrical signals from sensors measured by earth shakings, various indicators from medicine and hospital machines, like ECG, or many other specific sensors, they all represent a continuous measurement of corresponding physical quantity. This kind of procedure produces a continuous time series.

Time Series Types Classification

There are many various time series classifications based on some specific criteria. The most significant dependencies are length of the time step, stationarity and memory. Depending on the distance between time series, recorded values data are classified into:

- equidistant time series
- non-equidistant time series

Equidistant time series are formed, when its values are recorded periodically or ocasionally with a constant period length. A lot of environmental or physical processes are described by this kind of time series. Non-equidistant time series is defines as those time series, which do not keep the constant distance between observations. Econometric indicators, like stock prices, weather Forcasting are not necessary performed within regular time intervals, they are regulated by a concrete supply and demand rates on the specific market. Therefore, this kind of series suitably exemplify a non-equidistant time series example.

According to the rate of dependency between newly observed values and its last values, time series are divided into:

- long memory time series
- short memory time series

Time series with large memory are those, for which the autocorrelation function (ACF) decreases slowly. This kind of time series generally describes processes, which don't have fast turnovers. Traffic congestion, electric energy consumption, different physical or meteorological indicators, like air temperature measurements, all these processes are usually described by long memory time series. Short memory time series are those, for which autocorrelation function (ACF) is decreasing more rapidly. Typical examples which contain these processes from the econometric sector.

Another classification of time series is depend on their stationarity:

- stationary time series
- non-stationary time series

Stationary time series are those time series, for which statistical properties like variance or mean value, are constant over time. These time series rely in relative equilibrium in relation to its corresponding mean values. Other time series is belonging to non-stationary time series. In industry, economy or trading, time series more frequently belongs to the non-stationary category. In order to deal with the forecasting task, non-stationary time series are usually transformed to the stationary ones, by the appropriate pre-processing methods.

Time Series Components

Usually, most of the time series analysis methods assume, that time series data contains the systematic component (typically comprising several components) and random noise (error), which make difficult in detection of the regular components. Therefore, the majority of methods, includes different noise filtration methods, in order to detect the regular components, or it has to performed during data pre-processing. Most of the components belongs to two main classes. They belong to either a seasonal or trend component. The trend is a systematic linear or non-linear component, which can change over time. Seasonal component is a component which periodically repeating component. Both types of regular components are usually presented in the time series simultaneously. For example, sales may increase from year to year, but there is a seasonal component, which reflects the significant growth of sales in the month of December and a drop in August.

It's already mentioned, that the model of time series usually contains several components: trend component T(t), seasonal component S(t), random noise component R(t), and sometimes there is additionally mentioned a cyclical component C(t). The difference between cyclical and seasonal components is, that seasonal components represents a regular seasonal periodicity, while cyclical component has a longer lasting effect and may vary from cycle to cycle. Very often, the cyclical component is integrated into one trend component T(t). Fig. 5 demonstrate an example of time series decomposition.

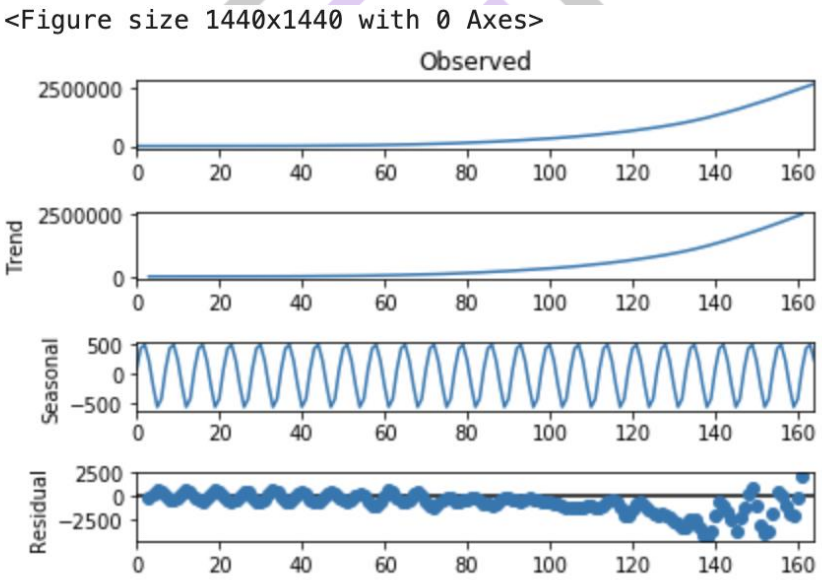

Figure 5.Time series components

Autocorrelation and Partial Autocorrelation

Dependencies between the actual and past values represent a fundamental principle of time series forecasting. It can be easily observed, that each value of the series is very similar to its neighbouring values. Additionally, time series contain a seasonal component, what means, that each value is also dependent on the values of similar time, but one season ago. Formally, any statistical dependency between two values is represent as a correlation, and is expressed by a corresponding coefficient..

Autocorrelation function

Autocorrelation (ACF) function calculates the correlations between the time series and its shifted copies at different points in time. The autocorrelations are usually calculated for the specific range of lags (shifts) and are expressed in the form of graph, called correlogram. Investigation of autocorrelation (ACF) enables to identify important dependencies int time series data (Covid-19 dataset.

Partial Autocorrelation Function

Sometimes it can happen, that the _first value is heavily dependent on the second value, the second value is heavily dependent on the third value and therefore the _first value is also dependent on the third value, and so on. This causes, that significant dependencies can be not found on the graph of autocorrelation (ACF) function. Partial autocorrelation function is another important tool. It is a modification of autocorrelation function (ACF), which allows to eliminate the described problem. Fig. 2.4 demonstrates results of autocorrelation (ACF) function and partial autocorrelation (PACF) function for the time series data from the previous section (Daily confermed Covid cases passenger counts from Jan-20 to August-20).

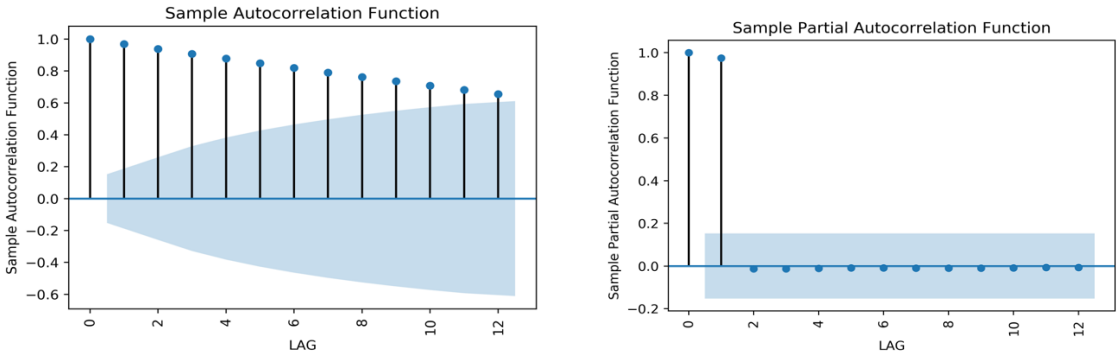

Figure 6.ACF and PACF of monthly Covid cases counts

V. FORECASTING METHODS

Forecasting is a technique to predicting future values through modelling the past and present data with the help of analyzing the season and trends. A common example might be an evaluation of some variable of interest at some specified future date. Prediction is a same, but more general term. Both introduce the general statistical methods engage the time series, cross-sectional data, or longitudinal data or alternatively to less formal judgmental methods. Usage can vary between areas of application: for example, in whether forecasting the terms "forecasting" are sometimes quite for estimates of values at certain specific future times, while the term "prediction" is used for general estimates, such as the forecasting the confirmed cases of Covid-19..

Auto Regression (AR) Model

The autoregression models (AR) the next step in the sequence as a linear function of the observations at prior time steps. The notation for the model involves specifying the order of the model p as a parameter to the AR function, e.g., AR(a). For example, AR (1) is a first order AR model.

Moving Average (MA)

The Moving average (MA) models is a model which is the next step in the sequence as a linear function of the residual errors from a mean process at prior time steps. A moving average (MA) model is a model that differs from calculating the moving average of the time series. The notation for the model involves specifying the order of the model c as a parameter to the MA function, e.g., MA(c). For example, Moving average MA(1) is a first-order moving average model.

Autoregressive Integrated Moving Average (ARIMA)

The Autoregressive Integrated Moving Average (ARIMA) models is a model which is the next step in the sequence as a linear function of the differenced observations and residual errors at prior time steps. It combines both Moving Average (MA) and Auto Regression (AR) models as well as a differencing pre-processing step of the sequence to make the series stationary, called integration (I). The notation for the model involves specifying the order for the AR(a), I(b), and MA(c) models as parameters to an ARIMA function, e.g., ARIMA (a, b, c). An ARIMA model can also be used to develop Autoregressive (AR), Moving Average (MA), and ARMA models.

Forecasting Model Comparison

Forecast Model Comparison

Forecasting Model and Method	Advantages	Disadvantages
Regression models	The main advantages of the given models are: simplicity, exibility and uniformity of calculations. Simplicity of model construction (only linear models).Transparency of all intermediate calculations.	Inefficiency and low adaptability of linear regression models for non-linear processes. Very complex non-linear model construction for the tasks with nonlinear functional dependency.
Autoregressive and Moving average models	Transparency and uniformity of calculations and model's construction. Relatively not complicated model construction. The most popular frequently used forecasting method. A lot of publications and information tells about how to implement this method for the specific problems.	A Large number of values required to be determined. Linearity, low adaptability and in-efficiency with non-linear processes.
Artificial Neural Networks models	The main advantage of these models is a non-linearity. Neural networks can easily deal with the non-linear dependencies between future and past values of the processes. Great adaptability and	Large number of parameters and significant options necessary to be selected. High hardware performance requirements during the network training process.it show architecture complexity and absence of transparency.

	scalability. Ability of parallel computations.	
Exponential smoothing models and methods	Transparency of intermediate calculations, simplicity and relative effectiveness. Easy model construction.	The disadvantage of this model is inexibility.

VI. DATA FORECASTING

The main tasks of this thesis are: analysis of provided time series data sets and development of the corresponding forecasting models. The dataset represents the time series of the number of confirmed cases of contagion reported by each country every day since the pandemic started. This dataset is available on the site of github of the Johns Hopkins University. All data sets that have been provided to me, contain Confirmed cases of COVID-19 infection are collected for India as well as other countries. Prediction of this kind of confirmed corona cases represents a real practical task and plays an important role for their further application..

In order to solve this problem, the practical part of thesis has been performed into two steps:

Data Analysis and Pre-Processing

The initial and the most important step in time series analysis is the determination of a time series dataset is stationary or not. This step is the initial step of the forecasting methods can deal only with stationary time series. In the theoretical part it was described, that stationary time series is one, whose statistical properties like mean and variance do not depend on time, at which the series is observed. From the practical point of view, the time series non-stationarity is usually caused by the presence of the seasonality or trend components inside the series..

Very often, simple visual observation of graphs of rolling variance and rolling mean functions that helps to make suggestion, whether the series is stationary or not. Both ACF and PACF belong to "rolling" analysis of the time series, when the sliding window technique is used to plot the progress of statistical parameters for the given size of window. The plot of rolling mean and rolling variance functions show the obvious evidence of the series non-stationarity. Definitely, the mean property do not show a constant progress over time, as well as the variance show the regular fluctuations.

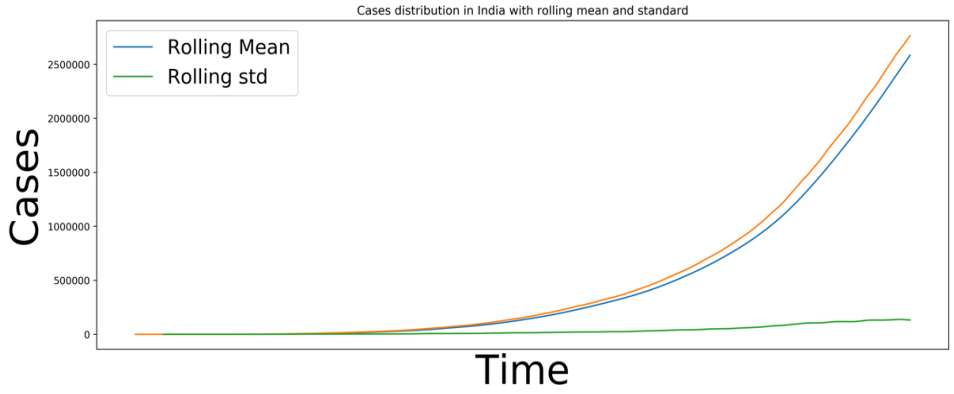

Figure 7. PACF of monthly Covid cases counts

Autocorrelation function

Autocorrelation (ACF) function calculates the correlations between the time series and its shifted copies at different points in time. The autocorrelations are usually calculated for the specific range of lags (shifts) and are expressed in the form of graph, called correlogram. Study of autocorrelation (ACF) enables to detect important dependencies in time series data.

Partial Autocorrelation Function

Sometimes it can happen, that the _first value is heavily dependent on the second value, the second value is heavily dependent on the third value and therefore the _first value is also depending on on the third, and so on. This causes, that significant dependencies can be not found on the graph of autocorrelation (ACF) function. Partial autocorrelation function is another important tool. It is a modification of autocorrelation function (ACF), which allows to eliminate the described problem. Fig. 2.4 demonstrates results of autocorrelation (ACF) function and partial autocorrelation (PACF) function for the time series data from the previous section (Daily confermed Covid cases passenger counts from Jan-20 to August-20).

Autoregressive Integrated Moving Average (ARIMA)

The forecasting models (ARIMA), that will be developed are based on the analysis performed in the section. In the beginning, the data set is divided into two parts training dataset 85% and testing dataset 15%. The next step is to analyze the results of Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) plots and make suggestions about the autoregressive and moving average parameters of the model. There are existing some general rules, how to identify the parameters [16]. The analysis in the last section demonstrated the necessity of one seasonal and one non-seasonal differencing. This suggests the use of ARIMA(a; b; c)_(A;B;C) model, where both differencing parameters b and B are equal to 1.

ARIMA Model Results

Dep. Variable:	D2.Cases	No. Observations:	169
Model:	ARIMA(5, 2, 4)	Log Likelihood	-1315.299
Method:	css-mle	S.D. of innovations	571.488
Date:	Sat, 26 Sep 2020	AIC	2652.598
Time:	19:17:54	BIC	2687.027
Sample:	02-01-2020 - 07-18-2020	HQIC	2666.570

	coef	std err	z	P>|z|	[0.025	0.975]
const	280.2642	132.122	2.121	0.034	21.310	539.219
ar.L1.D2.Cases	1.6626	0.090	18.422	0.000	1.486	1.839
ar.L2.D2.Cases	-1.8815	0.156	-12.097	0.000	-2.186	-1.577
ar.L3.D2.Cases	1.5724	0.197	7.987	0.000	1.187	1.958
ar.L4.D2.Cases	-1.1123	0.161	-6.906	0.000	-1.428	-0.797
ar.L5.D2.Cases	0.5613	0.079	7.125	0.000	0.407	0.716
ma.L1.D2.Cases	-1.7298	0.093	-18.593	0.000	-1.912	-1.547
ma.L2.D2.Cases	1.9635	0.141	13.909	0.000	1.687	2.240
ma.L3.D2.Cases	-1.3303	0.141	-9.455	0.000	-1.606	-1.055
ma.L4.D2.Cases	0.7010	0.094	7.445	0.000	0.516	0.886

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1508	-0.0000j	1.1508	-0.0000
AR.2	0.6618	-0.8399j	1.0693	-0.1438
AR.3	0.6618	+0.8399j	1.0693	0.1438
AR.4	-0.2464	-1.1372j	1.1636	-0.2840
AR.5	-0.2464	+1.1372j	1.1636	0.2840
MA.1	0.8289	-0.6185j	1.0342	-0.1020
MA.2	0.8289	+0.6185j	1.0342	0.1020
MA.3	0.1199	-1.1486j	1.1548	-0.2334
MA.4	0.1199	+1.1486j	1.1548	0.2334

Figure 8. Arima Model Result

This communicate to the use of one non-seasonal and one seasonal differencing. The remaining of parameters are going to be selected based on the analysis of ACF and PACF. In the Fig. 6 there can be observed, that plot of ACF tails o_ after the lag 11 and the plot of PACF tails of after the lag 9. This suggests, that autoregressive parameter a has to be tested up to 9 and moving average parameter c has to be checked up to 11. In the plot of ACF there can be also observed significant negative peak at lag 24, what corresponds to the effect of seasonal component. According to the referenced rules, this should be solved by adding seasonal moving average parameter to model. Each model will be trained on training dataset and then its performance will be tested on test dataset. To choose the optimal ARIMA model, the MSE rate will be used. Experimentally it has been tested, that ARIMA(5; 2; 4) model performs better than other models. The corresponding MAPE rate is 0.13%. Further increasing of the parameters(a, b, c) was pointless and didn't lead to improvement of performance. This is can be explained by effect of overfitting. Fig. 8 demonstrates the continuous plots of 1 day ahead forecasted values and the actual values of the test dataset.

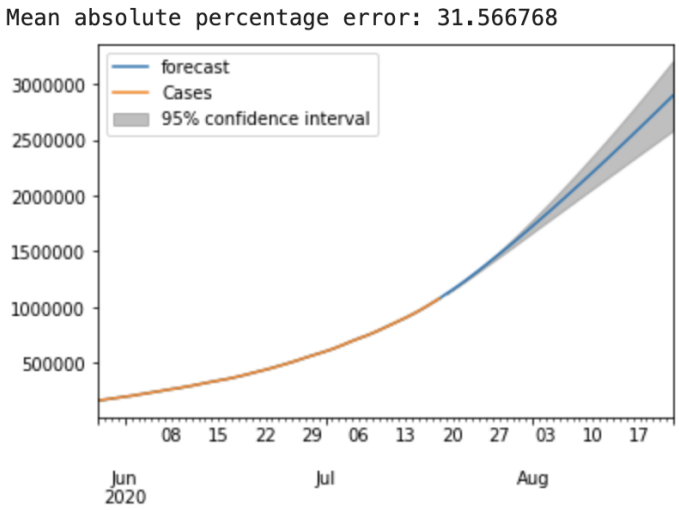

Figure 9. .The plots of one day ahead forecasted values (blue) and actual values

VII. Conclusion

The issues of this thesis is to perform the analysis of provided dataset and to develop the forecasting models for them. In order to solve this, the theoretical part of the thesis has been devoted to the survey of the time series problematic, forecasting methods, data pre-processing and other important aspects of time series analysis. It was investigated, that very often, the proper data pre-processing plays the key role of the whole process.

In the practical part of this thesis, there have been selected one perspective forecasting methods (ARIMA). Their effectiveness has been demonstratively tested on the Covid-19 data set, that is publicly available in the internet. Individual forecasting methods, as well as the time series analysis methods like ARIMA, SARIMA, have proven themselves in the experiments, by demonstration of

the remarkable results. After that, these methods could be confidently used for solving the main task of the thesis like forecasting models.

The main task of the thesis is related to the forecasting of the Covid-19 cases of individual country. Building of the qualitative forecasting models (ARIMA) for the coronavirus cases represents a real practical task and plays an important role for their further integration into the other applications. Forecasting performance of the so-called basic forecasting method", that simply adjusts the meteorological forecast and is currently used, has served as the standard values for the newly developed models.

Initially, for the covid forecasting, there have been used traditional forecasting methods without adding external factors. Afterwards, the models, that make it possible, have been extend by adding an external parameter. The meteorological day forecast values have served as the external factor. All forecasting methods demonstrated relatively good results, and better than the referenced benchmark value.

REFERENCES

[1] Gregory C. Reinsel Greta M. Ljung George E. P. Box, Gwilym M. Jenkins. Time Series Analysis: Forecasting and Control. Wiley, _fth edition, aug 2015.

[2] Michael Falk. A First Course on Time Series Analysis | Examples with SAS.Chair of Statistics, University of Wurzburg, aug 2012.

[3] James J. Filliben. Autocorrelation.
URL: http://www.itl.nist.gov/div898/handbook/eda/section3/eda35c.htm.

[4] Brian Borchers. The partial autocorrelation function, april 2001.
URL: http://www.ees.nmt.edu/outside/courses/GEOP505/Docs/pac.pdf.

[5] Jaakko Astola Ronald K. Pearson, Yrjo Neuvo and Moncef Gabbouj. Generalizedhampel _lters, may 2017.
URL https://tutcris.tut.fi/portal/files/7991765/Generalized_Hampel_Filters.pdf.

[6] Deepthi Cheboli. Anomaly detection of time series, may 2010.

[7] Simon S. HAYKIN. Neural networks and learning machines. New York: Prentice Hall,, third edition, 2009.

[8] University of Maryland prof. Charles Stangor. The neuron is the building block of the nervous system, may 2017.
URL http://2012books.lardbucket.org/books/ beginning-psychology/s07-01-the-neuron-is-the-building-blo.html.

[9] MSc Apostolos Panagiotopoulos. Optimising time series forecasts through linear programming, december 2011.
URL http://eprints.nottingham.ac.uk/12515/1/Apostolos_Panagiotopoulos_Thesis.pdf.

[10] Petros Kritharas. Developing a sarimax model for monthly wind speed forecasting in the uk, 2013.
URL https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/16350/2/Thesis-2014-Kritharas.pdf.

[11] Commonly used activation functions, may 2017. URL: http://cs231n.github.io/neural-networks-1/.

[12] Robert Nau. Statistical forecasting: notes on regression and time series analysis,may 2017.
URL http://people.duke.edu/~rnau/411home.htm.

[13] Robert Nau. The logarithm transformation, may 2017.
URL:http://people.duke.edu/~rnau/411log.htm.

[14] Robert Nau. Statistical forecasting: notes on regression and time series analysis,may 2017.
URL http://people.duke.edu/~rnau/411home.htm.