Forecasting new covid-19 cases using time series data

Project

Name - Amit Kumar
Reg. No. - 223510233787
Course - PGD Big Data Analytics

Superviser -
Dr. Pallavi Somvanshi
Associate Professor
School of Computational and Integrative Sciences
JNU New Delhi

School of Computational and Integrative Sciences
Jawaharlal Nehru University
New Delhi 110067

Table of Contents

Abstract

This project aims to develop an application for individuals planning travel to Indian subcontinent informing about the covid-19 situation in the country by forecasting new cases of COVID-19 using the ETS model on time series data. The dataset used in this study includes daily new cases of COVID-19 in the Indian subcontinent. The ETS model will be trained on this dataset to forecast the number of new cases. The results of this will provide estimated COVID-19 cases for next two weeks, the estimated number of new COVID-19 cases will help individuals make informed decisions about their travel plans to the subcontinent. This can contribute to the ongoing efforts to mitigate the impact of the pandemic on the the life of individuals.

Introduction

Covid-19

The COVID-19 pandemic has affected the entire world in unprecedented ways, causing widespread illness, death, and social and economic disruption. The disease is caused by the SARS-CoV-2 virus and is primarily spread through respiratory droplets when an infected person talks, coughs, or sneezes. COVID-19 can cause a range of symptoms, from mild to severe, and can be fatal in some cases, especially for older adults and those with underlying health conditions. The pandemic has forced governments and communities to take drastic measures to contain the spread of the virus, including lockdowns, travel restrictions, and vaccination drives. To prevent the spread of COVID-19, public health measures such as social distancing, wearing masks, and frequent hand washing are recommended. Vaccines have been developed and approved for use, which have been shown to significantly reduce the risk of severe illness, hospitalisation, and death. Despite these efforts, the virus continues to spread, and the pandemic has had a profound impact on public health, the global economy, and social interactions.

Indian subcontinent

The Indian subcontinent is a region in South Asia that includes the countries of India, Pakistan, Bangladesh, Nepal, Bhutan, Maldives and Sri Lanka. It is a densely populated and culturally diverse region with a rich history and complex geopolitics. The total population of the Indian subcontinent is approximately 1.8 billion people as of 2021. The Indian subcontinent is one of the most densely populated regions in the world. It is a rapidly developing region that includes travelling of millions individuals.

Map 1 : Map of Indian subcontinent

Forecasting

The COVID-19 pandemic has had a significant impact on the Indian subcontinent, affecting both the economy and society. The forecast of new COVID-19 cases can help to prepare healthcare systems and allocate resources effectively, implementing and adjusting mitigation strategies such as social distancing measures, travel restrictions, and vaccine distribution plans for minimising the impact of the pandemic on public health and the economy of the region.

It can also help individuals take necessary precautions to protect themselves and their loved ones from COVID-19. By understanding the potential trajectory of the pandemic, people can take appropriate measures such as wearing masks, practicing good hand hygiene, and avoiding large gatherings and travels.

Time series data

Time series data is a type of data that is collected over time at regular or irregular intervals. It involves a sequence of observations of a variable or set of variables taken at different points in time. Time series data can be continuous or discrete and can be recorded at various frequencies such as seconds, minutes, hours, days, weeks, months, or years.

Time series data can be decomposed into several components, which can help to better understand the patterns. These components include:

- Trend: The trend component represents the overall direction of the data over time. It can be upward, downward, or flat.
- Seasonality: The seasonality component represents regular, repeating patterns in the data that occur at fixed intervals, such as daily, weekly, or monthly.
- Cyclical: The cyclical component represents patterns in the data that occur over longer time periods, but are not necessarily regular or fixed.

Time series modelling is a powerful tool for analysing and forecasting, as it can capture the patterns and trends in the data and provide estimates for future trends.

About the data

Covid-19 data is collected from the WHO website containing daily data for each country with the following fields.

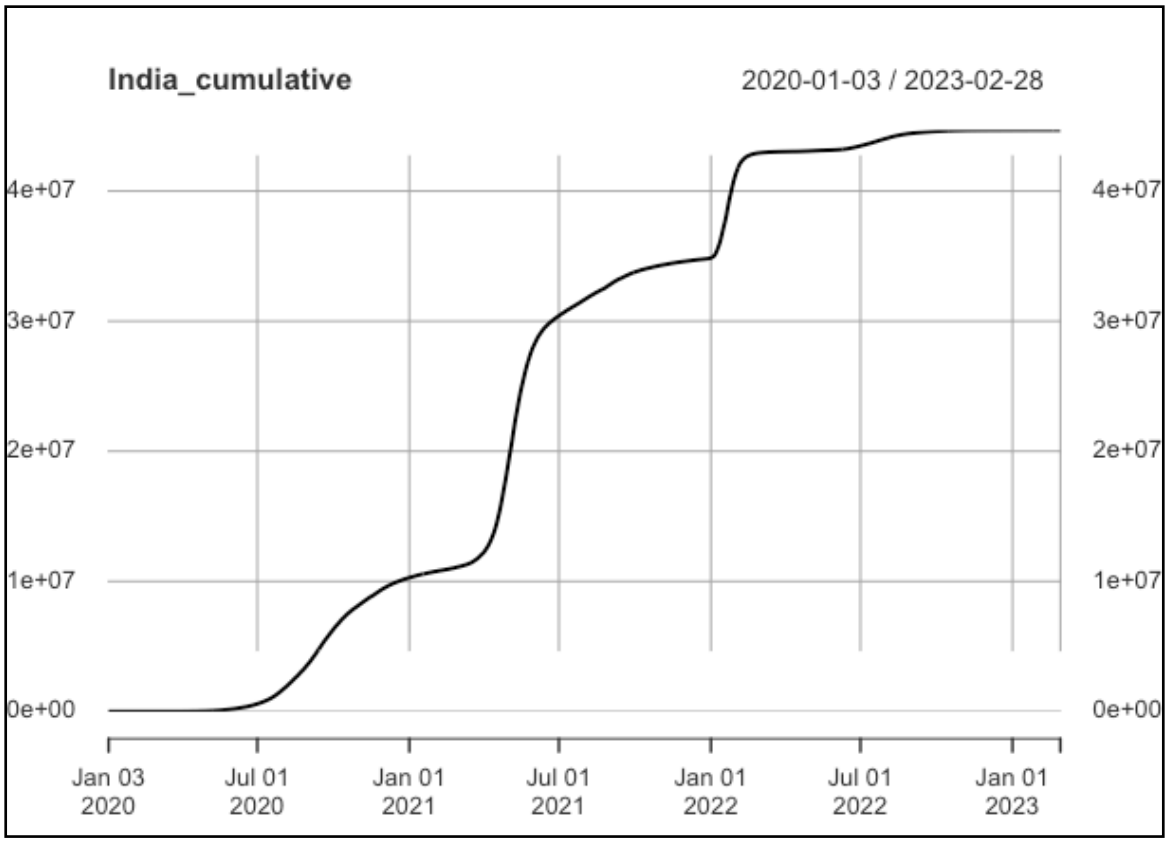
Field name	Type	Description
Date_reported	Date	Date of reporting to WHO
Country_code	String	ISO Alpha-2 country code
Country	String	Country, territory, area
WHO_region	String	WHO regional offices:
New_cases	Integer	New confirmed cases.
Cumulative_cases	Integer	Cumulative confirmed cases.
New_deaths	Integer	New confirmed deaths.
Cumulative_deaths	Integer	Cumulative confirmed deaths.

- As we are forecasting the covid cases for the Indian sub continent we only need data for 7 Indian sub continent countries.

Knowing our data

- The cumulative cases plot shows a consistent increase, it suggests the presence of a upward trend.

Graph 1: Cumulative cases of India



- These trends can make the data non-stationary by introducing a systematic change in the mean or variance of the series over time.
- To perform time series analysis and forecasting, it is important to have a stationary time series. A stationary time series is one whose statistical properties, such as mean, variance, and covariance, do not change over time. It is necessary to remove the trend component from the data. This can be achieved through various techniques.
- In our case the daily cases are given, which is already a stationary data.
- We are forecasting the new confirmed cases only so we only need the dates and new cases column.

Preprocessing of data

Console output 1 : Original Covid-19 data for India

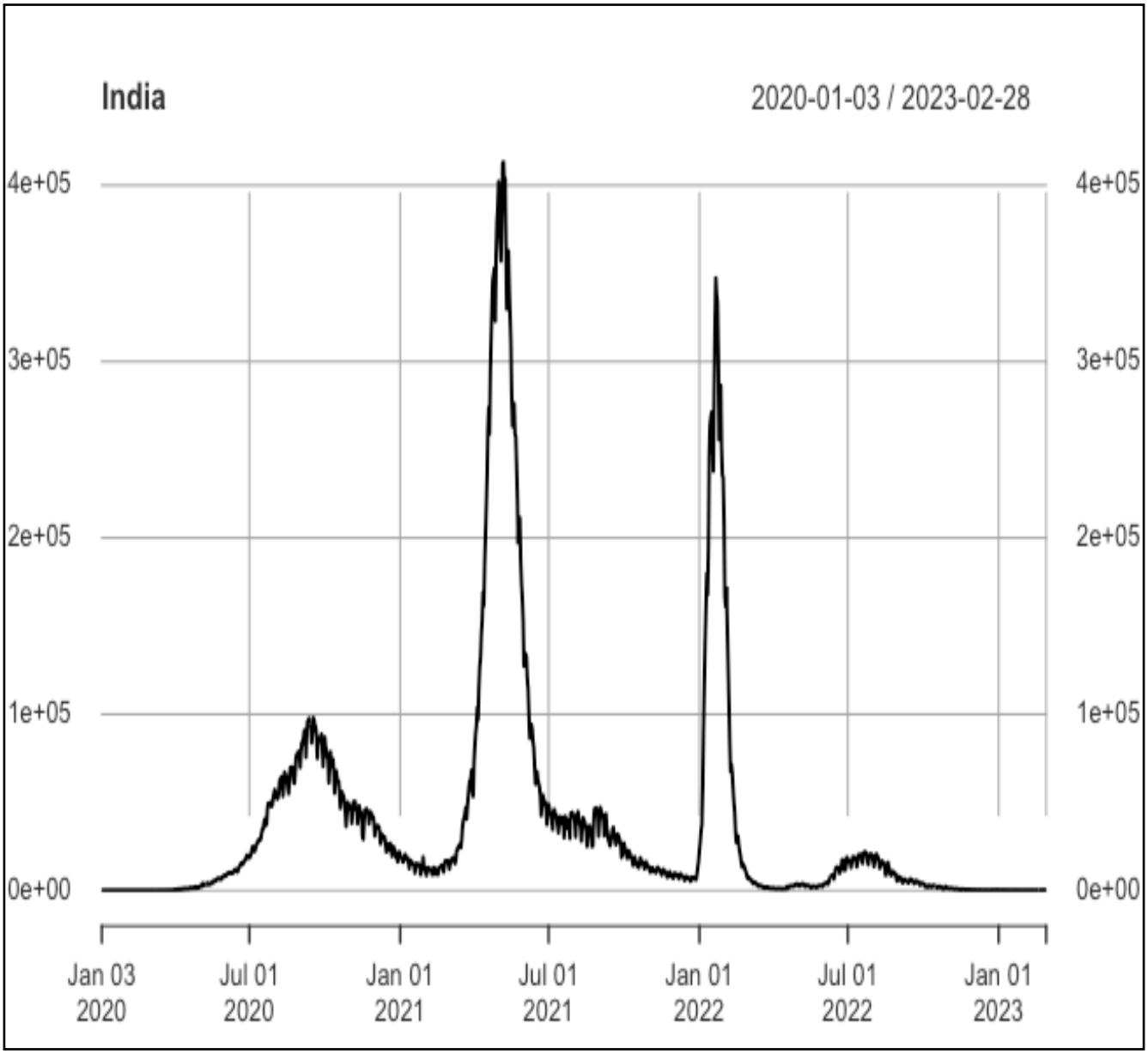
	Date_reported	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
111836	2023-02-23	IN	India	SEARO	193	44685450	1	530763
111837	2023-02-24	IN	India	SEARO	169	44685619	1	530764
111838	2023-02-25	IN	India	SEARO	180	44685799	0	530764
111839	2023-02-26	IN	India	SEARO	218	44686017	5	530769
111840	2023-02-27	IN	India	SEARO	185	44686202	1	530770
111841	2023-02-28	IN	India	SEARO	169	44686371	1	530771

- We are using time series forecasting model data, the data needs to be converted into time series data which involves organising the data into a sequence of values that are indexed by time that can be done via xts() function in r.
- About xts() function: -xts stands for "eXtensible Time Series,"provides a way to create time series data in R, with many useful features for analysing and visualising time series data.

Console Output 2 : After converting Covid-19 data into Time series data

	[,1]
2023-02-23	193
2023-02-24	169
2023-02-25	180
2023-02-26	218
2023-02-27	185
2023-02-28	169

Graph 2 : Plot for Daily covid-19 cases of India



Training model

Model used :- ETS

The ETS model is a statistical method used for time series forecasting. ETS stands for Error, Trend, and Seasonality, which are the three components that make up the model.

- Error - The random fluctuations or noise in the time series data that cannot be explained by the trend and seasonal patterns.
- Trend - The underlying long-term pattern, either increasing, decreasing, or stable.
- Seasonality - The repeating patterns that occur at fixed intervals.

The ETS model combines these three components to forecast future values of the time series. It uses past observations to estimate the parameters of the model, including the smoothing parameters that control the weights given to the different components. These parameters are optimised to minimise the sum of squared errors between the actual and predicted values of the time series.

Exponential smoothing is to forecast the next value of the time series based on a weighted average of past observations, where the weights decrease exponentially as the observations get older.

For simple exponential smoothing (SES), the formula is:

$F(t+1) = \alpha * Y(t) + (1 - \alpha) * F(t)$

where:

- F(t+1) is the forecast for the next time period
- Y(t) is the actual value of the time series at time t
- F(t) is the forecast for the current time period
- α is the smoothing parameter.

One of the advantages of the ETS model is its flexibility to handle different types of time series data, including those with trend, seasonal, and irregular components. It can also provide uncertainty estimates in the form of prediction intervals, which indicate the range of likely values for future observations. However, like any statistical model, the accuracy of the ETS model depends on the quality of the data and the appropriateness of the assumptions made about the underlying patterns in the time series.

Simulation of Model

For country India

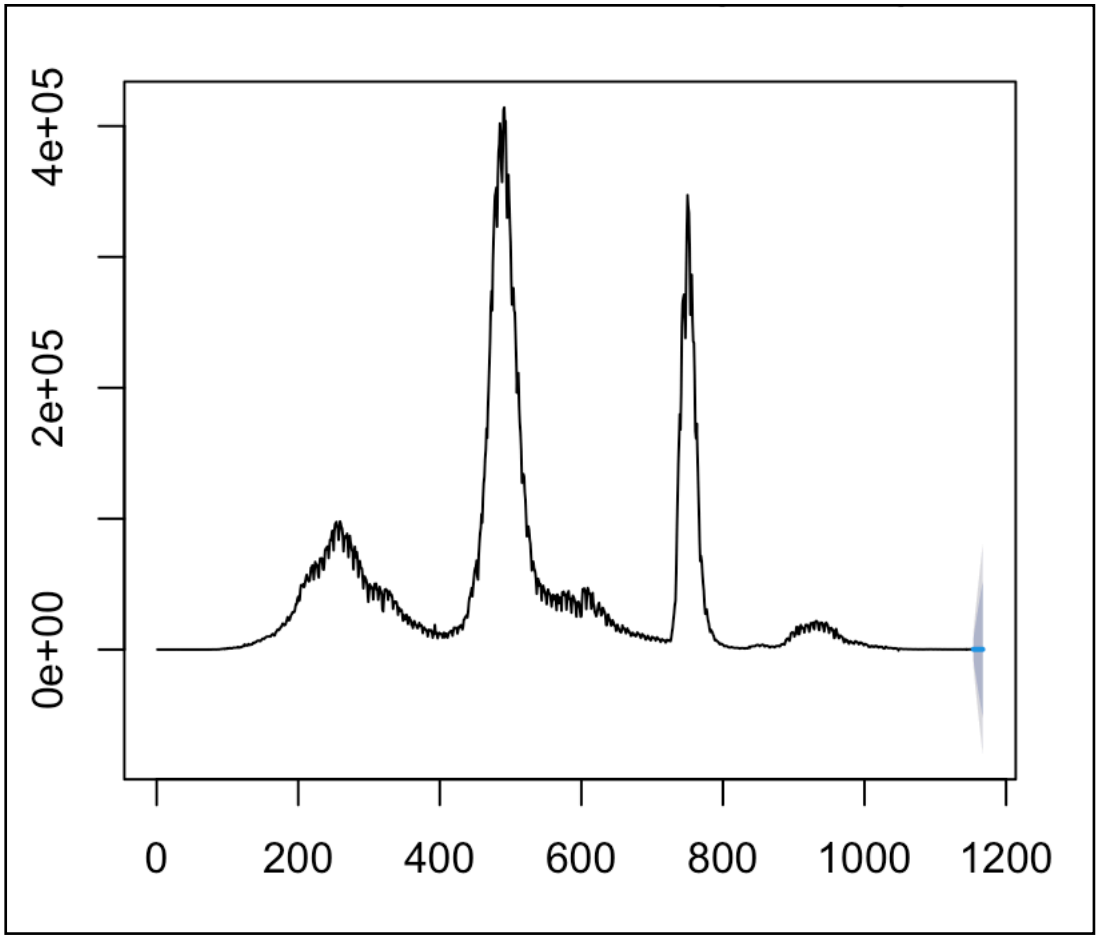
Console output 3 : Taking the input

```
1. India
2. Nepal
3. Pakistan
4. Bangladesh
5. Bhutan
6. Maldives
7. Sri Lanka
> country<-as.integer(readline("Enter the country you want to visit : "))
Enter the country you want to visit : 1
```

Printing the Expected minimum and maximum cases for next two weeks

```
Expected covid-19 daily cases for next two weeks would be between:  170  -  175
```

Console output 4 : Plot for forecasted cases for next 14 days



Error

To evaluate the accuracy of a forecasting model, four random periods were selected, and the forecasted cases were compared with the actual cases for each of those periods.

To determine the degree of accuracy, the Mean Absolute Percentage Error (MAPE) was calculated.

MAPE stands for Mean Absolute Percentage Error. It is a metric used to measure the accuracy of a forecasting model.

MAPE = (1/n) * ∑(|actual - forecast| / actual) * 100%

MAPE is expressed as a percentage, and the lower the value, the better the accuracy of the forecasting model.

Date - 06/02/2021 - 19/02/2021 **MAPE**=11.75

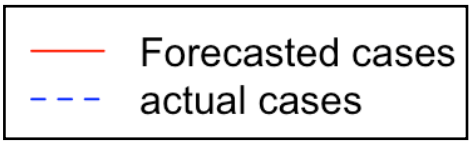
Date - 17/05/2021 - 30/05/2021 **MAPE**=27.96

Date - 03/12/2021 - 16/12/2021 **MAPE**=28.13

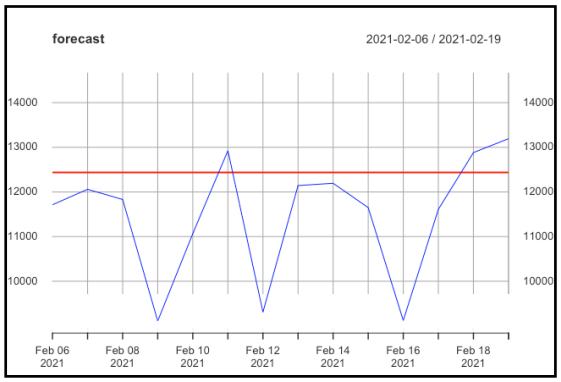
Date - 21/06/2022 - 04/07/2022 **MAPE**=13.17

Average MAPE = 20.25

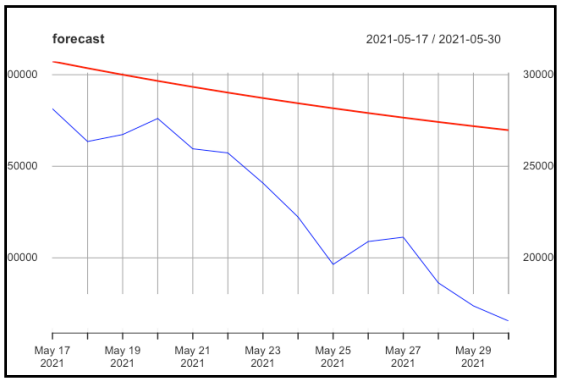
forecasted cases were compared with the actual cases



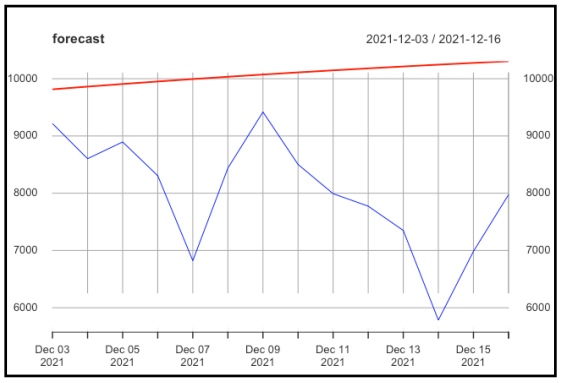
MAPE=11.75



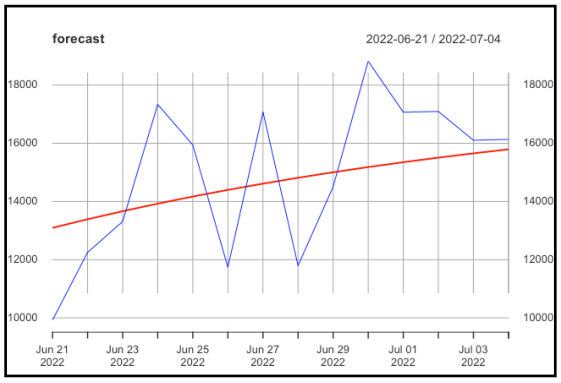
MAPE=27.96



MAPE=28.13



MAPE=13.17



Observations

1. Analysing the data, we have observed that there is an upward trend in the data.
2. we have found no seasonality in the data. This means that there is no regular pattern or cycle in the data that is repeated over a specific period.
3. However, we have identified the presence of cycles in the data. These cycles may be irregular and not follow a specific pattern or time period, but they do indicate that there are fluctuations in the values that occur over time.
4. Value of α determined by the model is always high(>0.8) that means our forecasted values are highly dependent on recent values.
5. The Mean Absolute Percentage Error (MAPE) of our forecast is around 20%. This indicates that there is a significant difference between our predicted values and the actual values. Given this level of inaccuracy, our forecast may not be highly reliable and should be viewed with caution.

Limitations

1. ETS models are sensitive to outliers in the data, which can distort the estimated parameters and lead to inaccurate forecasts.
2. ETS models are typically used for short-term forecasting, which is forecasting up to a few periods ahead. The reason for this is that ETS models are based on the assumption that the underlying patterns in the data are relatively stable over time. Therefore, they may not be appropriate for long-term forecasting, as the underlying patterns in the data may change significantly over longer periods of time.
3. In the context of COVID-19, where highly populated cities and urban areas can experience higher transmission rates than less populated rural areas.the data of a few major cities can have a significant impact on the overall data of a country, this can skew the overall data for the entire country.
4. Focusing solely on time series data. By avoiding other relevant factors such as density, spatial data, vaccine availability, literacy rate, age, and medical facilities, we may miss out on critical insights

Scope for improvements

1. Other relevant factors can be considered.
2. More advanced models can be used.
3. Forecasting for small clusters done.
4. Forecasting can be done for any place or any time.
5. Proper dashboard can be created to visualise the data better.

References

(1) Tableau (2022). Time Series Analysis: Definition, Types, Techniques, and When It's Used. [online] Tableau. Available at: https://www.tableau.com/learn/articles/time-series-analysis.

(2) Cowpertwait, P.S.P. and Metcalfe, A.V. (2009). Introductory time series with R. Dordrecht: Springer.

(3) Pankratz, A. (1991). Forecasting with dynamic regression models. New York, N.Y.: J. Wiley And Sons.

(4) Brownlee, J. (2017). How to Decompose Time Series Data into Trend and Seasonality. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/.

(5) rstudio-pubs-static.s3.amazonaws.com. (n.d.). Manipulating Time Series Data in R with xts & zoo. [online] Available at: https://rstudio-pubs-static.s3.amazonaws.com/288218_117e183e74964557a5da4fc5902fc671.html.

(6) Otexts.com. (2020). 7.7 Forecasting with ETS models | Forecasting: Principles and Practice. [online] Available at: https://otexts.com/fpp2/ets-forecasting.html.

(7) Encora. (n.d.). Exponential Smoothing Methods for Time Series Forecasting. [online] Available at: https://www.encora.com/insights/exponential-smoothing-methods-for-time-series-forecasting.

(8) Opuszko, Marek. (2020). Analytics of the COVID-19 (Corona) Spread using R.

(9) 7.6 Estimation and model selection | Forecasting: Principles and Practice (2nd ed). (n.d.). [online] otexts.com. Available at: https://otexts.com/fpp2/estimation-and-model-selection.html.

(10) Martinez, E.Z., Aragon, D.C. and Nunes, A.A. (2020). Short-term forecasting of daily COVID-19 cases in Brazil by using the Holt's model. Revista da Sociedade Brasileira de Medicina Tropical, 53. doi:https://doi.org/10.1590/0037-8682-0283-2020.