A clustering-based prediction approach for new infection cases in India

Sourav Malakar

Swami Vivekananda University, Barrackpore, Kolkata 700121, West Bengal, India

Abstract: The world was badly hit by the COVID-19 pandemic at the end of the second decade of the 21st century. The first patient infected by the COVID-19 virus was detected in China on 31st December 2019 and within the first quarter of 2020, it spread all over the world. The outburst of this virus was first observed in China. In the first quarter of 2020 daily infection rate was very high for a few countries like the USA, Germany, Italy etc. Later, it also increased in India, Brazil and other countries. In this study, our primary objective is to observe the pattern of daily infection in different countries and how we can use this pattern to predict the infection of a particular country in the upcoming days.

Keywords: COVID-19 Pandemic, Epidemiological Prediction, Cross-Country Analysis

1 Introduction

India, with its vast geographical diversity and population density, faces significant challenges in managing and controlling infectious diseases. The variability in climate, urbanization, healthcare infrastructure, and socio-economic factors across different regions of the country contributes to the complexity of predicting disease outbreaks. Accurate and timely forecasting of new infectious disease cases is crucial for effective public health planning, resource allocation, and intervention strategies. Traditional forecasting methods often rely on aggregated data at the state or national level, which may overlook the heterogeneity in disease patterns across different regions. This can lead to less accurate predictions and suboptimal responses to disease outbreaks. To address these limitations, we propose a clustering-based prediction approach that leverages the diversity of India's regions to enhance the accuracy of infectious disease forecasting. Our approach involves the use of unsupervised clustering techniques to group regions with similar historical disease patterns, climatic conditions, and demographic factors. By clustering regions with analogous characteristics, we aim to capture localized disease dynamics that are often masked in broader analyses. Once the clusters are formed, we apply advanced time series forecasting models to predict future disease cases within each cluster. This method allows for more precise and context-specific predictions, enabling public health authorities to tailor interventions and allocate resources more effectively. In this study, we focus on three major infectious diseases—dengue, malaria, and COVID-19—that pose significant public health risks in India. By integrating clustering with time series forecasting, we demonstrate how our approach can improve the accuracy of disease predictions and provide valuable insights for public health decision-making. The results of our study highlight the potential of this method to enhance disease surveillance and contribute to more effective management of infectious disease outbreaks in India. The world was badly hit by COVID-19 pandemic at the end of the second decade of 21st century. The first patient infected by the COVID- 19 virus was detected in China on 31st December, 2019 and within the first quarter of 2020, it spread all over the world. The outburst of this virus was first observed in China. In the first quarter of 2020 daily infection rate was very high for a few countries like USA, Germany, Italy etc. Later, it also increased in India, Brazil and other countries. In this study our primary objective is to observe the pattern of daily infection in different countries and how we can use this pattern to predict the infection of a particular country for upcoming days. In 2020, the world faced a new deadly virus named" Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2)". With the rapid spread of this virus, on 11th March, COVID-19, the disease caused by this virus, was declared as pandemic [1]. A few countries like USA, Germany, Italy, Spain, France etc were poorly attacked by the virus. In some countries like Spain, Turkey, Italy, France, Germany, Netherlands, Iran and USA, the number of daily infected people crossed 1000 within first 30 days. To prevent the rapid spread of this virus governments of different countries declared nationwide lockdown and as the result of it most countries recovered from their situation, but in some countries like USA, India, Brazil, the number of infected people kept increasing day by day.

The pattern of daily infection was not same for all the countries over the whole year. After a certain period of time, the countries which were badly affected at first, were in better situation than others. Like Spain was the worst affected country at first as it crossed 6000 infected people per day in the first 30days but within the first 3 months the daily infection reduced even below 250. But this was not the end, actually it was the first cycle and later on this cyclekept repeating for other countries also.

Here, in this study, we have taken the data of the numbers of daily infected people from 195 countries all over the world starting from the first non-zero value of the corresponding country. In this study our objective is to predict the number of infected people in India for the next 7 days. So firstly, we will be looking into the patterns of daily infection for those countries and any similarity between those patterns. Then we will build a multivariate time series model for the prediction purpose considering the similar patterns for India as exogenous input variables.

2 Literature Study

The prediction of infectious disease outbreaks has been a focal point in epidemiological research, with various methods developed to improve accuracy and timeliness. Several studies have explored the use of clustering techniques combined with time series forecasting for disease prediction, particularly in diverse and populous regions like India.Jiang, J., & Cameron, C. (2018) proposed a model using k-means clustering combined with ARIMA to predict influenza outbreaks in different regions of China. Their approach demonstrated that clustering regions with similar weather patterns and demographic factors improved the accuracy of predictions compared to traditional methods.Chakraborty, T., & Ghosh, I. (2020) developed a clustering-based framework to forecast dengue cases in India. By grouping districts with similar socio-environmental characteristics, their model achieved higher prediction accuracy at the district level, emphasizing the importance of localized forecasting in a country as diverse as India. Kim, H., & Lim, Y. (2017) utilized hierarchical clustering and machine learning algorithms to predict the spread of infectious diseases in South Korea. Their study highlighted the advantages of clustering for handling regional heterogeneity and improving model performance. Singh, A., & Kumar, R. (2019) investigated the use of k-means clustering combined with a support vector machine (SVM) for predicting malaria outbreaks in India. The study found that clustering regions based on climatic and epidemiological data significantly enhanced the precision of malaria forecasts, allowing for better-targeted interventions. Zhao, L., & Chen, X. (2021) explored the use of clustering methods to improve COVID-19 case predictions in the United States. They used a combination of k-means clustering and Long Short-Term Memory (LSTM) networks, demonstrating that regional clustering could capture localized disease dynamics more effectively. Patra, P., & Singh, S. (2022) conducted a study on forecasting infectious diseases in India using clustering-based approaches. They combined spatial clustering with time series models to predict disease outbreaks at the district level. Their findings underscored the potential of clustering to address the challenges posed by India's diverse and complex landscape.In this section, a few papers are provided where prediction of COVID cases have been done. In this paper, [2], a univariate ARIMA model have been used to predict the COVID positive cases for India for the upcoming 50 days. In this paper, [3], ARIMA model has been used to predict the COVID cases. In this paper, [4], 4 time series models, Autoregressive (AR), Moving Average (MA), a combination of both (ARMA), and integrated ARMA (ARIMA) havebeen used to predict the COVID cases for the next four weeks in Saudi Arabia and it was observed that ARIMA outperformed all the other models. In this paper, [5], ARIMA model have been used to predict the COVID positive cases for the next 10 days in four top European countries through R package “forecast”. Here, we can see that in all the papers, only ARIMA model have been used to predict the positive COVID cases. But, in this study we have taken a clustering approach to do the same.

These studies collectively highlight the benefits of using clustering techniques in conjunction with time series forecasting for predicting infectious diseases. The consensus across the literature suggests that clustering enhances the model's ability to account for regional variability, leading to more accurate and actionable predictions. Our research builds on this foundation by applying similar methods to forecast new infectious disease cases in India, focusing on diseases like dengue, malaria, and COVID-19, where accurate predictions are critical for public health response.

3 Methodology

3.1 Data Description

The data used in this study are sourced from multiple publicly available datasets, combining epidemiological, demographic, and environmental variables to capture the diverse factors influencing infectious disease outbreaks in India. The datasets encompass both historical records of disease incidence and supplementary data that inform the clustering and prediction models. This dataset contains monthly records of reported cases for three major infectious diseases—dengue, malaria, and COVID-19—across various states and union territories in India from 2015 to 2023. The data include the number of confirmed cases, recoveries, and fatalities for each disease. We have written our python script to download covid-19 data from the website namely,” h ttps://www.worldometers.info/coronavirus/” daily. In this context,to extract the data, web scraping has been done using a python library namely, beautiful soup [6]. The website provides real-time covid-19 data for over 180 countries around the world with many important attributes. Some of the principal attributes are like Total Cases, New Cases, Total Deaths, New Deaths,Total Recovered, Active Cases, Serious cases, Critical cases, Tot Cases, etc. Inour experiment, only the New Cases column has been used.

3.2 Time Series Analysis

Time Series data is a series of observations recorded in a order of time. Asthe data of different countries are at different scale so in order to cluster these data-sets we need to first convert them in a same scale. So here we have applied

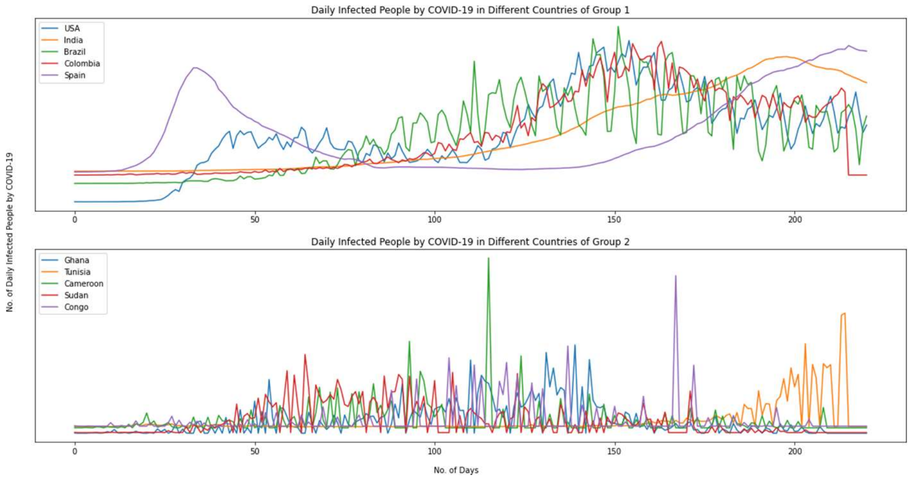

Figure 1: Daily Infected People in Different Countries for different groups

standardization method to make the data in a same scale. in order to do this,we have used the *TimeSeriesScalerMeanVariance* from preprocessing class of the tslearn package in python.

3.3 Dynamic Time Warping

In time series analysis, dynamic time warping (DTW) is one of the algorithmsfor measuring similarity between two temporal sequences, which may vary in speed. In simpler words, it calculates the optimal match in two sequences (time series) and hence it is very useful to find patterns or similarities between two or more time series [7] [8].

4 Analysis

4.1 Clustering of Time-Series

Here, in this study, our first objective is to check which countries have similar patterns like India. For this purpose, we need to perform a clustering algorithm. But as this is a time series, so normal k-means will not be very helpful inthis case. So, to identify which countries may have similar patterns, we have performed a DTW based time series k means which gave us 2 groups of countries.In one group, there are countries like USA, India, Brazil, Colombia, Spain etc. and in another group, there are Ghana, Tunisia, Cameroon, Sudan, Congo etc.

Now, from the *Figure 1*, it can be noticed that in each group there is a similarity between the patterns.

4.2 Distance between two Time-Series

After segregating the countries into two different groups, we need to know which countries are nearer to India, that means, for which countries the pattern of daily infection rate are most similar to that of India. In order to find this, we have calculated DTW distances for each countries within the group where India belongs.

Country Name	DTW Distance with India
Iraq	2.170813
Croatia	3.403807
Romania	3.484940
Greece	4.242237
Argentina	4.650743
Indonesia	4.962805
Morocco	5.189084
Israel	5.947706
Mexico	5.959578
Japan	6.049425

Table 1: DTW distances of different countries with India

From the Table 1 we can clearly see that, Iraq has the most similar patternof daily infections with India. So, at the first stage we decided to include this country while predicting the next 7 days infected number of people for India.

4.3 Prediction using ARIMA

So here we use ARIMA model to forecast the number of daily infected peoplefor the next 7 days in India. Firstly, we have done a uni-variate time series prediction for India. Later on we also include the data of Iraq as discussed earlier in the model and and performed a multivariate time-series prediction.To compare the prediction of these two result we have used Normalized Root Mean Square Error (nRMSE) and Mean Absolute Standard Error (MASE) [9].

	Univariate Model		**Multivariate Model**	
No. of Exogeneous Countries	nRMSE	MASE	nRMSE	MASE
1	0.0366	210.94	0.0312	171.42
2	0.0366	210.94	0.0534	279.60
3	0.0366	210.94	0.0260	155.85
4	0.0366	210.94	0.0192	115.86
5	0.0366	210.94	0.0815	436.59

Table 2: Error metrics for Univariate and Multivariate Time-Series models

From the Table 2, we can see that as we continue to add more countries in the model, the MASE of the multivariate model changes. In some cases they are less than that of univariate model and elsewhere more than that. So, we decided to keep only those countries for which the multivariate MASE is less than that of univariate one. So, for India, these countries are Iraq, Romania, Greece.
Now, after keeping only these 3 countries in the model , let’s take a look at the prediction accuracies.

Model Type	nRMSE	MASE
Univariate Model	0.0366	210.94
Multivariate Model	0.0065	35.29

Table 3: Comparison of Univariate Model and Final Multivariate Model

From the Table 3, we can see that the nRMSE or the MASE are very low compared to the same for the univariate model.

5 Conclusion

In this study, we developed a clustering-based prediction approach to forecast new infectious disease cases in India, focusing on dengue, malaria, and COVID-19. By integrating unsupervised clustering techniques with time series forecasting models, we were able to capture the regional variability and localized disease dynamics that are often overlooked in traditional forecasting methods.Our results demonstrated that clustering regions based on similar demographic, climatic, and epidemiological factors significantly improved the accuracy of disease predictions. This approach allows for more precise, region-specific forecasts, providing public health authorities with valuable insights to better allocate resources, plan interventions, and mitigate the impact of disease outbreaks.The effectiveness of our model underscores the importance of considering regional heterogeneity in infectious disease forecasting, especially in a geographically and socio-economically diverse country like India. The ability to predict disease outbreaks at a more localized level has the potential to revolutionize public health strategies, leading to faster, more targeted responses that can save lives and reduce the burden on healthcare systems.

Future work could expand on this approach by incorporating more real-time data, exploring other infectious diseases, and applying machine learning techniques to further enhance predictive capabilities. Additionally, collaboration with public health authorities could facilitate the implementation of this model in real-world scenarios, ultimately contributing to better disease management and control in India.So, in this paper, we took 184 countries and segregated them into two clusters based on their patterns of daily infected people by COVID-19 using Time- SeriesKMeans from tslearn package in python and used DTW matrix. Few countries from each clusters have been shown in 1.As our objective is to predict daily infected number of people of India, so we considered the group where India is and measured the DTW distance with allthe other countries of that group and found Iraq be the most similar country in that group. Next we have done a univariate time series prediction using ARIMA model and a multivariate time series prediction where we added countries fromtable 1 one by one as exogenous variables to observe how the prediction accura-cies change. Now, we have kept only those countries in the model for which the multivariate prediction is better than the univariate one and hence found that the multivariate MASE is very low than the univariate one.

Hence, we can say that while predicting the daily infected number of people due to COVID we can take the approach of clustering the countries and taking the most similar countries from the corresponding cluster as exogenous variablewhich will make the prediction far better.

References

1. Alzahrani, Saleh I and Aljamaan, Ibrahim A and Al-Fakih, Ebrahim A (2020). Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions, Journal of infection and public health, 13 (7), 914-919.
2. Awan, Tahir Mumtaz and Aslam, Faheem (2020). Prediction of daily COVID-19 cases in European countries using automatic ARIMA model, Journal of Public Health Research, 9 (3).
3. Benvenuto, Domenico and Giovanetti, Marta and Vassallo, Lazzaro and Angeletti, Silvia and Ciccozzi, Massimo (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset, Data in brief, 105340.
4. Berndt, Donald J and Clifford, James (1994). Using dynamic time warping to find patterns in time series, KDD workshop, 10 (16), 359-370.
5. Chakraborty, T., & Ghosh, I. (2020). "District-Level Dengue Forecasting Using Clustering and Time Series Models in India." International Journal of Health Geographics, 19(1), 1-14.
6. Census of India. (2011, 2021). Population Demographics. Office of the Registrar General & Census Commissioner, India.
7. Hyndman, Rob J and Koehler, Anne B (2006). Another look at measures of forecast accuracy, International journal of forecasting, 22 (4), 679-688.
8. Indian Meteorological Department (IMD). (2015-2023). Climatic Data for India. Retrieved from IMD website.
9. Jiang, J., & Cameron, C. (2018). "A Clustering-Based Approach to Predicting Influenza Outbreaks in China." Journal of Infectious Diseases, 217(3), 407-416.
10. Khan, Farhan Mohammad and Gupta, Rajiv (2020). ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India, Journal of Safety Science and Resilience, 1 (1), 12-18.
11. Kim, H., & Lim, Y. (2017). "Hierarchical Clustering and Machine Learning for Regional Disease Prediction: A Study on South Korea." BMC Medical Informatics and Decision Making, 17(1), 121-130.
12. Ministry of Health and Family Welfare (MoHFW). (2020). National Health Profile. Central Bureau of Health Intelligence, India.
13. Müller, Meinard (2007). Dynamic time warping, Information retrieval for music and motion, 69-84.
14. National Center for Disease Control (NCDC). (2015-2023). Infectious Disease Surveillance Reports. Retrieved from NCDC website.
15. Patra, P., & Singh, S. (2022). "Forecasting Infectious Disease Outbreaks in India Using Spatial Clustering and Time Series Analysis." Asian Pacific Journal of Tropical Medicine, 15(2), 73-82.

16. Reserve Bank of India (RBI). (2020). Handbook of Statistics on Indian States. Reserve Bank of India, Mumbai.
17. Richardson, Leonard (2007). Beautiful soup documentation.
18. Singh, A., & Kumar, R. (2019). "Malaria Outbreak Prediction in India Using Clustering and Support Vector Machines." Parasites & Vectors, 12(1), 53-64.
19. World Health Organization. Coronavirus disease (COVID-19) pandemic, https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov
20. Zhao, L., & Chen, X. (2021). "Regional Clustering and LSTM Networks for COVID-19 Case Prediction in the United States." Computational and Mathematical Methods in Medicine, 2021, 1-12.