International Journal of E-Health and Medical Communications
Volume 13 • Issue 2 • July-December 2022

Time-Series Forecasting and Analysis of COVID-19 Outbreak in Highly Populated Countries:
A Data-Driven Approach

Arunkumar P. M., Karpagam College of Engineering, Coimbatore, India
Lakshmana Kumar Ramasamy, Hindusthan College of Engineering and Technology, Coimbatore, India
Amala Jayanthi M., Kumaraguru College of Technology, Coimbatore, India

ABSTRACT

A novel corona virus, COVID-19, is spreading across different countries in an alarming proportion, and it has become a major threat to the existence of human community. With more than eight lakh death count within a very short span of seven months, this deadly virus has affected more than 24 million people across 213 countries and territories around the world. Time-series analysis, modeling, and forecasting are important research areas that explore the hidden insights from larger set of time-bound data for arriving at better decisions. In this work, data analysis on COVID-19 dataset is performed by comparing the top six populated countries in the world. The data used for the evaluation is taken for a time period from 22nd January 2020 to 23rd August 2020. A novel time-series forecasting approach based on auto-regressive integrated moving average (ARIMA) model is also proposed. The results will help the researchers from the medical and scientific communities to gauge the trend of the disease spread and improvise containment strategies accordingly.

KEYWORDS

ARIMA, COVID-19, Data Analysis, Disease Spread, Time-Series Forecasting

1. INTRODUCTION

The emergence of novel corona virus is identified from the Wuhan City, Hubei province in China during December 2019 and subsequently renamed as COVID-19 by World health organization. The most common symptoms of the virus include fever, cough and tiredness. Some lesser known symptoms are headache, diarrhea, sore throat and loss of taste or smell. Most of the severe cases of COVID-19 showed symptoms of breathing difficulty and chest pain. Monitoring of epidemiological changes in an in-depth manner will give better perceptions on the disease outbreak (Rotha & Byrareddy, 2020). The research on time-series data is highly critical due to the enormous usage of temporal data in wide variety of applications. Large dataset, high dimensionality and frequent updation are few characteristics of time-series data. The time-series data is subjected to various processing steps to discover the patterns for better decision making. Apart from pattern discovery and clustering, other important task of time-series data mining include classification, rule mining and summarization (Fu,2011). Distance-based clustering, fuzzy c-means (FCM) algorithm, Autoregressive integrated

DOI: 10.4018/IJEHMC.20220701.oa3

moving average (ARIMA) models and Hidden Markov model (HMM) are few methods adopted for time-series clustering and pattern discovery. Time series forecasting depends on the task of analyzing past observations of a random variable and generates a model that portrays the underlying relationship and its patterns. Each of the forecasting method follows four important steps namely, problem definition, information gathering, selecting the best model and forecasting (Hyndman & Athanasopoulos,2018). The time-series analysis and forecasting for COVID-19 disease outbreak is an emerging research paradigm that requires deep knowledge and better experimentations for interpreting the trend and evaluating the predictions.

Holt–Winters Additive Model (HWAAS), Auto-regressive integrated moving average (ARIMA), TBAT, Prophet, DeepAR and N-Beats and Vector Auto regression (VAR) are few models used by researchers around the world for time-series forecasting(Papastefanopoulos,2020). In HWAAS model, trend and seasonal variation of the data are taken in to account. This method is an advanced model proposed by adopting added features to Holt's exponential smoothing. In exponential smoothing, the recently recorded observations are used for updating the prediction levels. The additive method is favored when the seasonal variations are approximately constant through the data series. Holt-Winters Exponential Smoothing is also called as Triple Exponential Smoothing. TBAT method involves four components, namely, Trigonometric seasonal formulation, Box–Cox transformation, ARMA errors and trend component (Harvey et al., 1997; Box & Cox,1964; Adhikari & Agrawal,2013). Multiple seasonalities can be accommodated by TBAT model. Here, each seasonality is modeled with a trigonometric representation based on fourier series. Prophet method is proposed by Facebook. Three major components used by this model are trend, seasonality and holidays.

Time-series decomposition of Prophet model is given by the equation as:

$$Y(t) = A(t) + B(t) + C(T) + \hat{e}\,(t) \tag{1}$$

where, A(t), B(t) and C(T) denotes trend, seasonality and holiday. The last term ê (t) implies error value. Saturating growth model and a piece-wise linear model are utilized by Prophet approach to gather forecasting results. DeepAR deploys a long short-term memory based recurrent neural network architecture for time-series forecasting. Probabilistic forecasting is supported by this model that trains an auto regressive recurrent network (Salinas et al., 2020). N-Beats model is the short form of Neural basis expansion analysis for interpretable time series forecasting. A deep neural architecture consisting of forward and backward residual links is used by N-Beats model (Oreshkin et al.,2019). In this method, generic architecture and an interpretable architecture are used in tandem and dual residual stacking results are observed. Vector Auto regression (VAR) model is a simplification of the univariate autoregressive model for forecasting a vector of time series. It is a multivariate forecasting algorithm .Such model should possess at least two time series variables that influence each other. In ARIMA model, information in the past values of the time series can alone be used to predict the future values.

2. MATERIALS AND METHODS

The COVID-19 data is retrieved from Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The data contains cumulative count of infected people observed in daily basis. The data is fetched for a time period from 22nd January 2020 to 23rd August 2020. The work is segregated in twofold. The first part of the work demonstrates the data analysis of COVID-19 impact in six highly populated nations. For this work, six months of time-series COVID-19 data (starting from 22-01-2020) is used for evaluation. The second part of the work utilizes ARIMA model for time-series prediction of COVID-19 disease by analyzing two nations, namely, India and US. The COVID-19 dataset consists of daily time series summary of confirmed, recovered and mortality