Chapter 3
Predicting Daily Confirmed COVID-19 Cases in India:
Time Series Analysis (ARIMA)

Sudip Singh
Uttar Pradesh Rajya Vidyut Utpadan Nigam, India

ABSTRACT

India, with a population of over 1.38 billion, is facing high number of daily COVID-19 confirmed cases. In this chapter, the authors have applied ARIMA model (auto-regressive integrated moving average) to predict daily confirmed COVID-19 cases in India. Detailed univariate time series analysis was conducted on daily confirmed data from 19.03.2020 to 28.07.2020, and the predictions from the model were satisfactory with root mean square error (RSME) of 7,103. Data for this study was obtained from various reliable sources, including the Ministry of Health and Family Welfare (MoHFW) and http://covid19india.org/. The model identified was ARIMA(1,1,1) based on time series decomposition, autocorrelation function (ACF), and partial autocorrelation function (PACF).

INTRODUCTION

Coronavirus disease (COVID-19) is an infectious disease caused by a Novel Coronavirus discovered in 2019. COVID-19 virus is primarily transmitted through droplets discharged when an infected person coughs, sneezes, or exhales. India has seen significant increase in daily confirmed cases despite initiating stringent lockdown measures across the country.

Time Series is a collection of data points collected at constant time intervals. These are analyzed to determine the long term trend so as to forecast the future or perform some other form of analysis. Since covid-19 spreads from person to person, the basic assumption of a linear regression model that the observations are independent doesn't hold making a case for time series analysis. Covid-19 daily confirmed data series has an increasing trend, while seasonality trends may also exist, i.e. variations specific to a particular time frame.

DOI: 10.4018/978-1-7998-7188-0.ch003

Various authors have applied ARIMA and other time series models to predict cases for other diseases like Ebola virus disease in west African countries (Manikandan et al., 2016), incidence of dengue in Rajasthan (Bhatnagar et al., 2012), hemorrhagic fever with renal syndrome in China (Liu et al., 2011). Zhang X, Zhang T, Young AA, Li X provided for comparisons of four time series models in epidemiological surveillance data in his paper (Zhang et al., 2014). In the context of covid-19 cases in India, Santanu Roy, Gouri Sankar Bhunia, Pravat Kumar Shit have used Geographic Information System based approach to apply ARIMA(2,2,2) and predicted which Indian district are highly vulnerable for COVID-2019 (Roy & Bhunia, n.d.).

In this paper we have applied simple prediction of daily Covid-19 confirmed cases in India. In further paper we will apply other time series methodologies like Seasonal ARIMA (SARIMA), Vector Auto Regression (VAR) and Vector Error Correction Model (VECM) to predict Daily Confirmed, Active and Recovered Cases respectively and compare the predictions with ARIMA(1,1,1).

ARIMA MODEL

A Time Series is said to be stationary if its statistical properties such as mean, variance remain constant over time. Time Series models are based on the assumption that the Time Series is stationary. Various techniques like differencing, log-differencing etc. are employed to make a non-stationary Time Series stationary. ARIMA model is a combination of differencing with auto-regression and a moving average model. ARIMA is an acronym for Auto-Regressive Integrated Moving Average (in this context, "integration" is the reverse of differencing).

We call this a non-Seasonal ARIMA(p, d, q) model, where

'p' is the order of the AR term. it refers to the number of Y lags which should be used as predictors.
'd' is the number of differencing required to make the time series stationary. It is the minimum amount of differentiation needed to render the sequence stationary, and if the time series is stationary already, then $d = 0$.
'q' is the order of the MA term. it refers to the number of lagged errors in the forecast that should go into the ARIMA model.

The objective of this paper is to fit a non-stationary ARIMA model to correctly recognize the stochastic mechanism of the time series of Covid-19 Daily Confirmed cases and forecast future values for the same.

METHODOLOGY

The best parameters of the ARIMA Model (suitable lags for the components of the AR and MA and the number of differencing required to induce stationarity) is determined for the Daily Confirmed Covid-19 Time Series data. The Auto Correlation (ACF) function and the Partial Auto Correlation (PACF) function are used to determine the best model.

Simple plot of the Daily Confirmed Covid-19 Time Series data is given (Figure 1). It is visually evident from the lag plot of Daily Confirmed Covid-19 Time Series data that there is strong correlation (Figure 2). Further, we observe some significant correlation between t=1 and t=28 (roughly) with significant