ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.6, 3947–3953
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.6.75 Spec Issiue on ICAML-2020

ARIMA MODELLING FOR TIME SERIES DISEASE FORECASTING

J. DEV[1] AND N. SHARMA

ABSTRACT. This manuscript presents the time series forecasting modeling using Auto Regressive Integrated Moving Average (ARIMA) for communicable disease forecasting. In this regard, the disease dataset related to Dengue and Tuberculosis (TB) is collected for the past 17 years i.e., the year 2002 to 2009. Different ARIMA models are applied to check the efficacy of the ARIMA model for disease forecasting. The advanced prediction for Dengue and Tuberculosis is required to initiate a pro-active program in advance in order to restrict the spreading of the disease and eventually minimize the loss of human life. Beforehand prediction can also lead to the eradication of such diseases in the long run. In this case, Time Series Forecasting can be used to deal with this and predict the future outcomes of these diseases. ARIMA Model is put to use in the time series for a comparative analysis of the diseases mentioned above.

1. INTRODUCTION

Time series forecasting involves the use of a numerical model to find out the future values of the time series, on the basis of the data provided in the past, [1,2]. It contains data that can be traced and accumulated with the passage of time. In these applications, the data is collected throughout the passage of time, and this data is hence used by the time series to predict the future values. The aim of data learning is either to learn about the observed environment or to extract information that encourages decision-making, [3].

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 62P10, 60G25.

Key words and phrases. ARIMA, forecasting, time series, communicable diseases, prediction.

In the future, time series forecasting could become an essential tool for a healthy world, [4]. Societies around the world are bound to have older populations and higher rates of non-communicable diseases, needless to spell the ongoing COVID 19 crisis. Tools like time series forecasting shall facilitate the medical fraternity tools to facilitate patients from the initial period, putting them onto treatments that are shown by robust, real-world data to lead to the best outcomes for their health issue.

Further, Dengue is a viral disease caused by female mosquitoes that has started to spread quite rapidly in various countries in the past few years. It is of the species, Aedes aegypti. Although it is less common, sometimes people may develop severe Dengue, which can cause complications like severe bleeding, plasma leakage, and organ impairment. Severe Dengue thus can prove to be quite fatal, [5].

Furthermore, TB or Tuberculosis is another deadly disease caused by a bacterium named "Mycobacterium tuberculosis", which most often affects the lungs. It is curable and preventable, [6]. TB is spread from one person to another through the air. Thus, when an infected person sneezes, coughs, or spits, he/she propel TB germs into the air, and the person who inhales them might have the chance of catching this disease. People with lower immunity might have a higher risk of getting affected by it. People with active TB can affect an average of 5-10 more people with this over the course of a year, [7]. Without proper preventive measures or treatment, it can prove to be highly lethal.

This manuscript outlines the ARIMA time series forecasting technique applicable to the medical domain.

2. Proposed Methodology

Proposed methodology is detailed as under elaborating the steps followed in the analysis of the data. Detailed explanation of the steps followed in building the suitable forecasting model is given below:

2.1. **Goal Definition.**

The first step of the data analysis starts with Goal Definition, which involves data collection, refinement, and analysis through various visualizing tools.

2.2. **Data Collection.** After carefully examining the data received after defining the goal of the forecasting that is to be done; the data is collected proceeding towards the time series forecasting process.

2.3. **Time Series Decomposition.** To choose the adequate components of time series forecasting, it is more convenient to divide the time series data into systematic part and a non-systematic part. The systematic components are divided into three parts: level, seasonality and trend.

- Level provides the average value of the series.
- Trend is the net change in the initial value and the final value of the data.
- Seasonality describes a short termed recurring behavior of data that can be seen multiple times in a series.

The non-systematic part of the Time Series is referred to as the Noise.

2.4. **Time Series Exploration and Visualization.** With the help of data visualization, we can find out initial patterns, identify its components and look out for some potential problems such as extreme values, unequal spacing, and missing values.

2.5. **Apply Forecasts.** At a given time t, the forecast which is k steps ahead is given by

$$F_{t+k} = y_t.$$

2.6. **Measuring Predictive Accuracy.** Predictive accuracy tell us how well the method has performed as compared to other measuring forecasts used as a reference level. Thus, to forecast the predictive functioning, several measures are generally used to judge the predictive accuracy of a forecasting method.

3. Results and Discussion

For implementation purpose, R Studio has been used as the programming software to forecast the occurring of diseases like Tuberculosis and Dengue. The data is taken from www.chp.gov.hk, [8]. This data tells us about the number of people who were affected by Dengue in Hong Kong (From 2002-2019), and

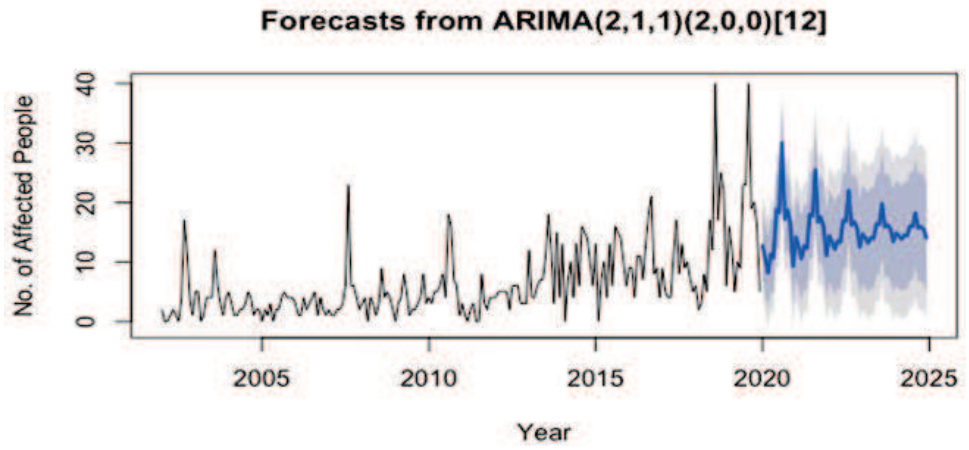

FIGURE 1. Modeling of Dengue Cases

the number of deaths that were caused due to Dengue (From 1995-2019) in the same area. For the case of Dengue disease, it is observed that the data has a monthly frequency, throughout the given time span. Auto ARIMA function is applied to it, to find the best fit forecasting model for the data. It was then forecasted and the data was plotted on a graph given in Figure 1.

In Figure 1, the black graph line shows the number the people who were affected by Dengue from 2002 till 2019. The blue line shows the projection of the number of people who might get affected by the disease in the upcoming five years, starting from 2019. From this we can infer that the number of people who will be affected by Dengue will be less than the year 2018 and 2019, and the count will continue to decrease in the coming years.

The data for tuberculosis was also taken from the same source. It has a frequency of 18 in a given year, starting from 1995 till 2019. The data is sorted on the basis of and provides inputs on number of deaths due to Tuberculosis. It is divided into three parts to give the forecast for males and females in Hong Kong, moreover, an overall forecast is also produced to study the overall trend of deaths caused by tuberculosis.

Figure 2, Figure 3, Figure 4 represent the number of people who were suffering from tuberculosis, because of which they died. The black line (From1995-Till 2019) reflect the number of people who died because of TB. The blue line gives us the most probable forecast of the number of people who might also die due to this disease in the upcoming five years. Since in all the figures the general death

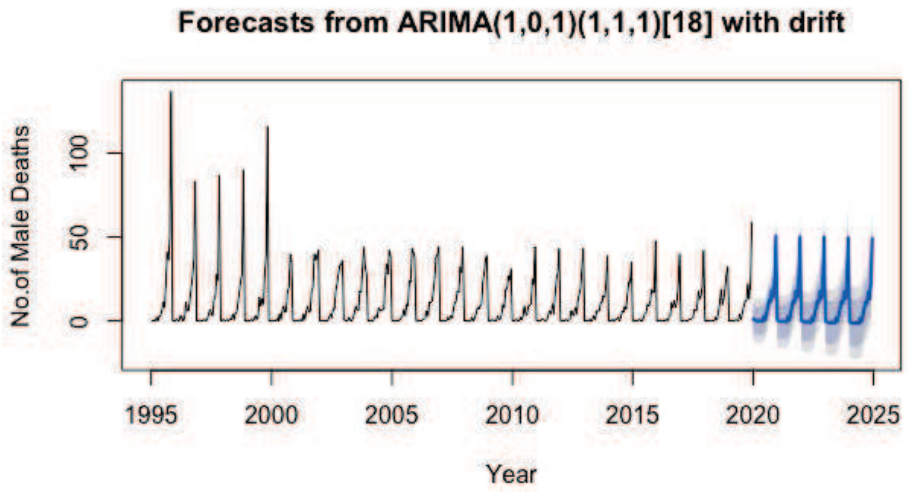

FIGURE 2. Forecasts for Tuberculosis cases for no. of Male Deaths

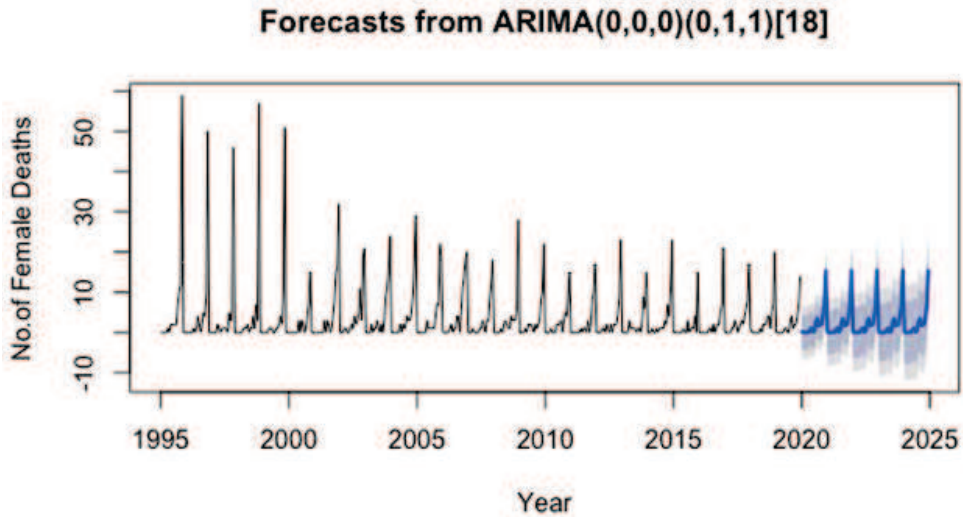

FIGURE 3. Forecasts for Tuberculosis cases for no. of Female Deaths

rate pattern has been quite similar since the year 2000, therefore the forecast tells us that somewhat similar pattern will be observed in the forthcoming years also.

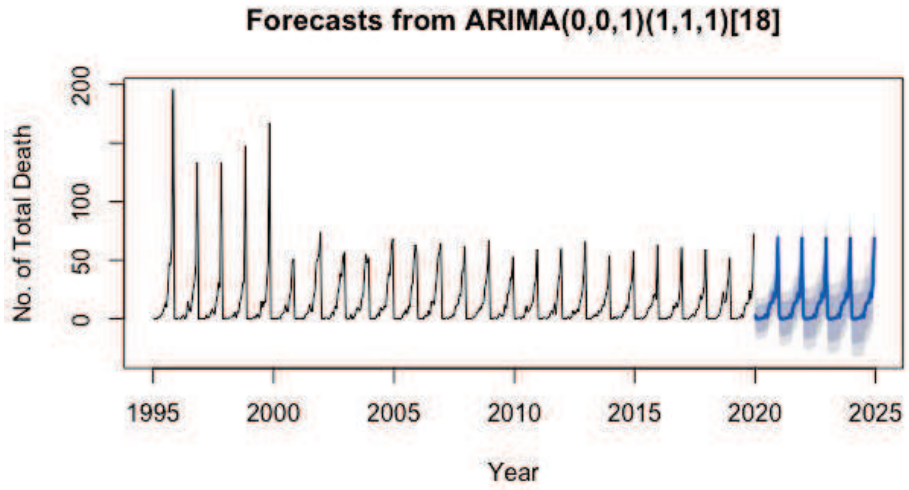

FIGURE 4. Forecasts for Tuberculosis cases for the Overall Trend

4. CONCLUSION

Time series forecasting and its related tools are a fast emerging technology and in the times to come shall combine the power of statistics, artificial intelligence, historical data available and the insights to create a predictive solution to the issues concerning the globe and its citizens. These tools shall facilitate a peep into the future, basis what has happened and is happening. Industries of all kinds shall undergo a sea change when this tool shall be used.

REFERENCES

[1] P. J. BROCKWELL, R. A. DAVIS: *Introduction to time series and forecasting*, Part of the Springer Texts in Statistics, Springer, Cham, 2016.

[2] V. ZARNOWITZ, A. OZYILDIRIM: *Time series decomposition and measurement of business cycles, trends and growth cycles*, Journal of Monetary Economics, Elsevier, **53**(7) (2006), 1717–1739.

[3] S. VERMA, N. SHARMA: *Statistical Models for Predicting Swine F1u Incidences in India*, First International Conference on Secure Cyber Computing and Communication (ICSCCC), (2018), 134–138.

[4] N. SULTANA, N. SHARMA: *Statistical Models for Predicting Swine F1u Incidences in India*, First International Conference on Secure Cyber Computing and Communication (ICSCCC), (2018), 134–138.

[5] Y. ZHANG, T. WANG, K. LIU, Y. XIA, Y. LU, Q. JING, J. LU: *Developing a time series predictive model for dengue in Zhongshan, China based on weather and Guangzhou dengue surveillance data* , PLoS neglected tropical diseases, **10**(2) (2016), e0004473.
[6] H. WANG, C. W. TIAN, W. M. WANG, X. M. LUO: *Time-series analysis of tuberculosis from 2005 to 2017 in China*, Epidemiology & Infection, **146**(8) (2018), 935–939.
[7] J. F. DE OLIVEIRA, T. B. LUDERMIR: *A hybrid evolutionary decomposition system for time series forecasting*, Neurocomputing, **180** (2018), 27–34.
[8] *www.chp.gov.in.*

MAYOOR SCHOOL
NOIDA, UTTAR PRADESH, INDIA
Email address: jaidityadev1402@gmail.com

DR. B.R. AMBEDKAR NATIONAL INSTITUTE OF TECHNOLOGY,
JALANDHAR, PUNJAB, INDIA.
Email address: nonita@nitj.ac.in