Grenze International Journal of Engineering and Technology, June Issue

Trend-Forecasting of Covid-19 Cases using Regression Anlaysis and Python

M. Lakshmana Rao[1], B Raghu Kumar[2], V.Kalyanamanohar[3] and G.Harinath Gowd[4]
[1]Mechanical Engineering Department, Madanapalle Institute of Technology & Science, Madanapalle, India
Email: raonaidupalem@gmail.com
[2]Mechanical Engineering Department, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada
Email: braghu5051@gmail.com
[3]Mechanical Engineering Department, University College of Engineering, JNTU, Kakinada
[4]Mechanical Enigeering Department, Vemu Institute of Technology
Email: gowdmits@gmail.com

***Abstract*— The objective of this paper is to compare the registered cases and predicted cases of covid-19 using regression analysis. The effect of covid-19 is more on society. If there is any 4th wave or 5th wave of covid, we would suffer even more. So, to prevent this disaster, Author has attempted in estimating number of covid cases. Here problem taken from the state government. Using forecasting techniques and compared predicted and real values. Then developed a code through python programming. Through this analysis, it has been observed that there is a good agreement between the analytical and programmed trend. Trend forecasting algorithm was developed through python coding.**

***Index Terms*— Trend, Forecast, Prediction, Algorithm, Python**

I. INTRODUCTION

Time series modelling is a dynamic research area which has attracted attentions of research community over last few decades. The main aim of time series modelling is to carefully collect and rigorously study the past observations of a time series to develop an appropriate model which describes the inherent structure of the series. This model is then used to generate future values for the series, i.e., to make forecasts. Time series forecasting thus can be termed as the act of predicting the future by understanding the past [1]. A time series is a sequential set of data points, measured typically over successive times. It is mathematically defined as a set of vectors x(t), t = 0,1,2 ... where t represents the time elapsed [2-4]. The variable x(t) is treated as a random variable. The measurements taken during an event in a time series are arranged in a proper chronological order. Times series is advantageous as it helps to (a) understand the past behavior (b) plan future operations and (c) facilitates comparison between past and future observations. A time series has four types of movement called elements of time series, namely, secular trend, seasonal variation, cyclical fluctuations, and irregular movements. Secular trend represents the movement that persists over a long period of time. The fluctuations, which complete the whole sequence of changes, within the span of a year and has about the same pattern year after year, is called a seasonal variation. Pronounced fluctuations moving up and down every few years throughout the length of the chart refer to business cycles or cyclical fluctuations. The little irregularities on any original t rend curve are referred to as irregular

Grenze ID: 01.GIJET.10.2.564_1

movements.

Time series observations are of high importance in many domains such as business, economics, industry, engineering, and science, etc. It is obvious that a successful time series forecasting depends on an appropriate model fitting. A lot of efforts have been done by researchers over many years for the development of efficient models to improve the forecasting accuracy. The present work aims to improve trend accuracy using least square method to forecast Covid-19 cases. Exponential moving average method is used for smoothening the obtained time series trend.

II. METHODOLOGY

A. Measurement of Trend-Forecasting

There are four methods used for determining the trend (rise), namely, free-hand or graphic method, semi-average metho, moving average method, and method of least square [5]. Out of these methods, least square is highly practical and mathematical method and with its help a trend line is fitted to the data in such a manner that the following conditions are satisfied.

(1) $\Sigma\,(Y - Yc) = 0$. This means the sum of deviations of the actual values of Y and the compound value of Y is Zero.

(2) $\Sigma\,(Y - Yc)^2$ is least. This means the sum of the squares of the deviations of the actual and compound values is least from this line. The line obtained by this method is known 'as the line of best fit'. The method of least squares is used to fit a straight-line trend. This straight-line trend is represented by the equation,

$$Yc = a + bX$$

where Yc is the trend value as distinguished from actual value, is Y values, a denotes the computed trend figure of the Y variable when X =0, and b is the slope of trend line.

In order to determine the values of the constants 'a' and 'b', the following two normal equations are to be solved.

$$\Sigma Y = Na + b\Sigma X \qquad \ldots \text{(i)}$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \qquad \ldots \text{(ii)}$$

where N is the number of years, months or period for which data are given

The above two normal equations would take the form of,

$$\Sigma Y = Na \qquad \ldots \text{(a)}$$

$$\Sigma XY = b\Sigma X^2 \qquad \ldots \text{(b)}$$

The values of a and b now be determined easily,

$$\text{Since} \quad \Sigma Y = Na \qquad \therefore a = \frac{Y}{N}$$

$$\text{Since} \quad \Sigma XY = b\Sigma X^2 \qquad \therefore b = \frac{XY}{X^2}$$

The constant 'a' is imply equal to the mean of Y values and the constant 'b' gives the rate of change

B. Trend Analysis

Trend analysis is carried out by plotting several points for the demand levels which is experienced in the past. After carefully examining these points draw a smooth curve _ through them and then predict the form (linear, exponential, or other) by free-hand or by statistical method that fits the historical data.

Trend Forecasting Algorithm Steps

1) Identify the Problem and Defining the problem
2) Collect Information from the source
3) Perform a Preliminary Analysis
4) Choose the Forecasting Model
5) Data analysis
6) Verify Model Performance

Smoothing Time Series (Moving Average)

When the actual sales for number of years is supplied (random nature of data), then it is not possible to describe this business time series by simple mathematical function since the data may not really show a simple linear, non-linear trends. In such cases, the time series can show a coherent form for forecasting purpose, when random variations in the data have been smoothed out by using moving average, which applies the same weightage to each of the data values in the averaging interval. It may be noted here that the degree of smoothing must depend on the number of periods forming the average interval. The short moving average allows much of the peaks and valley pattern of the original data to remain, while the longer moving average has a very much pronounced influence in smoothing out these variations.

Smoothing Time Series by Exponential Moving Average

The most known method of smoothing the effects of Random variations in the demand is the application of exponential weighted moving average. The exponential weighted moving average is an average that is based on geometric progression resulting in non-uniform weights for different data values of time series. The reason for doing so is to lay stress on the most recent data, and the weighted average can achieve this purpose.

The notable feature of this is that the maximum weight is given to the most recent information and the least to the oldest information. Comparing the two extremes, we see a gradation of weight attached to the data values. As the information becomes progressively older, the weightage becomes consequently smaller. Exponential smoothing retains all the desired features of the moving average, while eliminating undesirable aspects. It is not necessary to supply large amount of data. The sensitivity can be easily modified by changing the magnitude of 'SMOOTHING CONSTANT'.

III Results And Discussion

The data was collected from the government of Andhra Pradesh health, medical and family welfare's official website. The number of registered covid cases on each day between 11-04-2022 and 20-04-2022 is presented in Table 1.

Table I. Covid-19 Data Obtained From Govt's Website.

DD-MM-YYYY	Day	Y	X	XY	X^2
11-04-2022	1	24	-5	-120	25
12-04-2022	2	15	-4	-60	16
13-04-2022	3	19	-3	-67	9
14-04-2022	4	45	-2	-90	4
15-04-2022	5	47	-1	-47	1
16-04-2022	6	8	1	8	1
17-04-2022	7	38	2	76	4
18-04-2022	8	37	3	111	9
19-04-2022	9	44	4	176	16
20-04-2022	10	75	5	375	25
	n = 10	**ΣY = 352**		**ΣXY = 362**	**ΣX^2 = 110**

$$a = \frac{Y}{N} = \frac{352}{10} = 35.2 \quad b = \frac{XY}{X^2} = \frac{362}{110} = 3.29$$

$$\text{Now, } Yc = a + bX = 35.2 + 3.29\, X\, 6 = 54$$

The registered cases on previous day (20-04-2022) = 722.

Forecast for 21-04-2022 = 722 + 54 = 776.

Similarly, the forecast was made for each day until 26-04-2022 and the forecast values were compared with Govt's data, and a graph was plotted das shown in Figure 1. A good agreement between our forecast data and Govt's official data was found.

A plot was drawn using the data mentioned in Table 1 as shown in Figure 1(a). The scatterplot shows the reference line with registered cases scattered on its either side.

In the next stage, the trend-forecasting was done using python. The code was written in IDLE as mentioned below.

```
import matplotlib.pyplot as plt
import numpy as np
```

```
def estimate_coef(x, y):
number of observations/points
n = np.size(x)
# mean of x and y vector
m_x = np.mean(x)
m_y = np.mean(y)
# calculating cross-deviation and deviation about x
SS_xy = np.sum(y*x) - n*m_y*m_x
SS_xx = np.sum(x*x) - n*m_x*m_x
# calculating regression coefficients
b_1 = SS_xy / SS_xx
b_0 = m_y - b_1*m_x
return(b_0, b_1)
def plot_regression_line(x, y, b):
# plotting the actual points as scatter plot
plt.scatter(x, y, color = "m",marker = "o", s = 30)
# predicted response vector
y_pred = b[0] + b[1]*x
# plotting the regression line
plt.plot(x, y_pred, color = "g")
# putting labels
plt.xlabel('x')
plt.ylabel('y')
# function to show plot
plt.show()
def main():
# observations / data
x = np.array([-5,-4,-3,-2,-1,1,2,3,4,5])
y = np.array([24,15,19,45,47,8,38,37,44,75])
# estimating coefficients
b = estimate_coef(x, y)
print("Estimated coefficients:\nb_0 = {} \
\nb_1 = {}".format(b[0], b[1]))
# plotting regression line
plot_regression_line(x, y, b)
if __name__ == "__main__":
main()
```

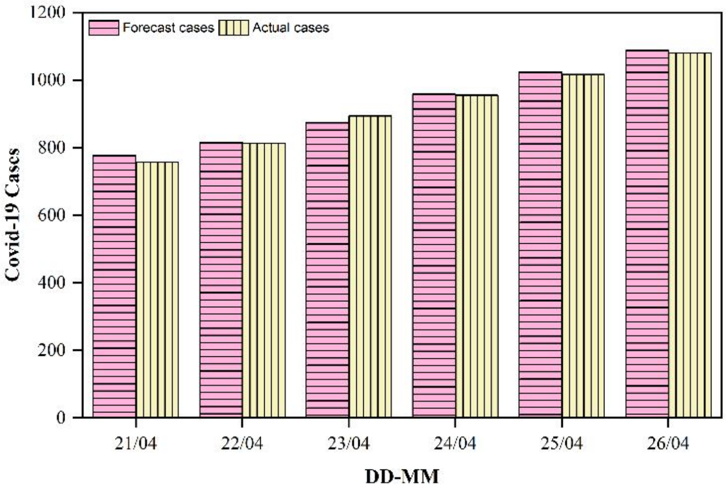
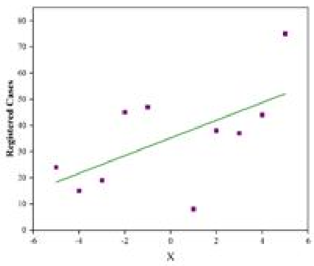
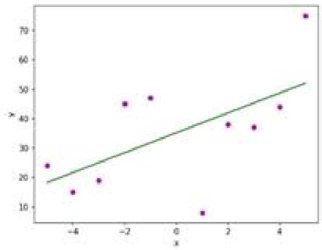


Figure 1. Forecast and actual covid-19 cases.



(a)



(b)

The output obtained from the graph is a plot as shown in Figure 1(b). It is clear that the graph plotted based on calculated output bears good resemblance to that plotted via python code. Thus, it is confirmed that the written code is accurate.

Smoothing time series

Smoothing time series was done to handle nonlinear variation in the Covid-19 data. Exponential moving average method was used for smoothing time series and the obtained results are presented in Table 2.

TABLE II. FORECASTING OF NON-LINEAR COVID-19 CASES.

DD-MM-YYYY	Actual Cases	Forecast Cases
14-04-2022	200	100
15-04-2022	50	120
16-04-2022	150	106
17-04-2022	145	115

A smoothing constant = 0.2 was used for the present analysis. Accuracy can be obtained by suitably adjusting the smoothing constant.

IV. CONCLUSION

The following conclusions are drawn from the current study.

- The trend-forecasting was measured using least square method.
- The trend-forecast was plotted using the covid data obtained from Government of Andhra Pradesh's official health ministry website.
- Python language was used develop a code to predict the trend-forecast algorithm. The code presented reasonably good agreement with the analytical trend.
- The python-based method presented in this study is beneficial for health industries for forecasting covid trend.

REFERENCES

[1] Raicharoen, Thanapant, Chidchanok Lursinsap, and Paron Sanguanbhokai. "Application of critical support vector machine to time series prediction." In Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS'03., vol. 5, pp. V-V. IEEE, 2003.

[2] Zhang, G. Peter. "A neural network ensemble method with jittered training data for time series forecasting." Information Sciences 177, no. 23 (2007): 5329-5346.

[3] Zhang, G. Peter. "Time series forecasting using a hybrid ARIMA and neural network model." Neurocomputing 50 (2003): 159-175.

[4] Tong, Howell. Threshold models in non-linear time series analysis. Vol. 21. Springer Science & Business Media, 2012.

[5] Chapman, Stephen N. The fundamentals of production planning and control. Upper Saddle River, NJ, USA: Pearson/Prentice Hall, 2006.