Time series forecast of Covid 19 Pandemic Using Auto Recurrent Linear Regression

DOI : 10.36909/jer.15425

Ferdin Joe John Joseph

Thai-Nichi International College, Thai-Nichi Institute of Technology, Bangkok, Thailand

ferdin@tni.ac.th

Abstract

Covid 19 pandemic has done severe impact in the economy and lifestyle of the people since the beginning of 2020. Various data analytics has been tried on the data obtained from various sources. These analytics include symptoms prediction, time series forecasting and impact analysis. The forecast on when the pandemic ends is a challenge for many countries. Time series forecasting models have been proposed for various applications but a non-seasonal and non-stationary forecasting method is needed to predict the progression of the pandemic. An Auto Regressive Linear Regression (ARLR) Algorithm is proposed in this paper with a selected geography's Covid data. The results of the proposed methodology sounds convincing when compared to the non-seasonal and non-stationary existing methodologies like linear regression and exponential smoothing variants. The performance measure of standard deviation and RMSE of the proposed method obtained 430.22 and 0.31 for active cases while 27.01 and 0.77 for rate of transmission with positive skew and platykurtic trend.

Keywords: Covid 19, Time Series Forecasting, Recurrent Linear Regression, Pandemic forecasting, Auto Recurrent Linear Regression

Introduction

Covid Pandemic has paralyzed the economy, health and lifestyle of people drastically in the first half of the year 2020. Anticipation is high among people on when the pandemic will come under control and lead a normal life. Covid 19 pandemic is due to the variant of Severe Acute Respiratory Syndrome (SARS) which is said to be originated from China in 2019 (Velavan and Meyer, 2020) . There are many mathematical models to predict the future trend of pandemic. The most important mathematical base is time series forecasted analysis. Usually time series forecast analysis is done for seasonal, stationary and exponential data. Pandemic data forms a different trend of seasonal data which is shown in the bell diagram mentioned in figure 1.

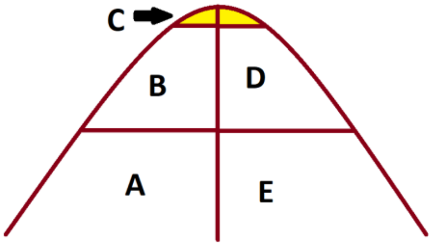

Figure 1: Active cases curve during a pandemic

The parts of A and B goes with a positive slope whereas D and E goes with a negative slope marking the rise and downfall of the active cases during the pandemic. Hump C has near zero slope which marks the peak of impact done by the pandemic.

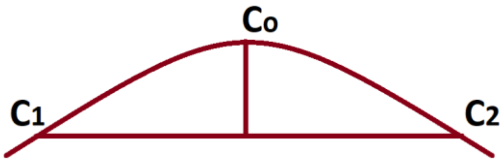

Figure 2: Hump C of pandemic curve

The tangent β_1 corresponding to the arc C_1C_o is the marking of near peak and the tangent β_2 corresponding to the arc C_oC_2 is the marking of the identifier which leads to the decrease of impact by the pandemic. A pandemic curve in a particular geography can have single or multiple humps and each hump has the parts mentioned in figures 1 and 2. These humps play an important role in predicting the trend of pandemic using time series forecasting. Cumulative cases over the pandemic curve follows an exponential trend but it is not helpful in predicting the time when the pandemic will come under control.

The proposed methodology is further explained in experimentation and the resultant of experiments are illustrated in the results and discussion section and concluded ultimately.

Related Work

Time series analysis and forecasting has been done for decades with huge corpus of mathematical formulations. Univariate discrete time series is a stochastic process with probabilistic terms. It is an autoregressive model that deals with the previous values in the series known as Auto Regressive Integrated Moving Average (ARIMA). An auto correlation function used by (Brockwell and Davis, 1996) for managing the municipal solid waste forecasting in Spain. The peaks observed in this method shows 12 lag units which perfectly matches with the seasonal trend. For a non-seasonal trend and stationary behavior (Takens, 1981) proposed a stationary method of Auto Regressive Moving Average (ARMA). There are also some nonlinear time series analysis methods using discrete map like (Gershenfeld and Weigend 1994). These methods are the basic and those proposed in the genesis of time series analysis. Time series analysis has been used in various applications. The scope of time series is discussed extensively [6] and the main applications are from stock prediction, electricity utility, business applications etc.

1. Medical data forecasting

Severity of illness from ICU data is predicted using multivariate time series modelling approach (Sapankeiwich and Sankar, 2009) uses Gaussian process. This was evaluated with the categorization of clinical assessment, time series abstraction and Gaussian methods. Daily active cases were predicted using multivariate SARIMA (Ghassemi, 2015) and this method is closely related to the active case prediction of Covid 19. Similarly in (Kam et al 2010), auto regressive and multinomial distribution of the number of calls was used to predict the actual incoming number of calls to the emergency service. Since the emergency health data is seasonal, the methodology for pandemic progression should be refined to go along with a backpropagation algorithm.

2. Covid Time series analysis

In correlation to the medical data based applications, Covid 19 pandemic progression is also discussed in a limited number of literatures. A stacked encoder (Hu et al, 2020) for transmission of epidemics was developed for data available in China during its receding phase. Average error was calculated to justify the effectiveness of the stacked encoder. ARIMA model on covid data is done (Benvenuto et al, 2020) and a correlellogram was presented on the method proposed. Rate of infection was statistically analyzed (Deb and Majumdar 2020) using the time dependence pattern and evaluated for the root mean square error between actual and predicted data. This domain specific time series cannot follow seasonal trend. A split based analytics was proposed in (Mizumoto and Chowell, 2020) which presented pandemic progression between three different classified groups of people using an estimate of mean reproduction number. Psychological aspects of the pandemic progression (Petropoulos and Makridakis, 2020) was presented with a timeline of events and the impact it had on the people. This was like a sampled quantitative study done on people. A deep learning LSTM method was proposed to predict the time series of pandemic in Canada is presented in (Yang et al 2020) and evaluated using RMSE. Most of the methodologies are evaluated using RMSE, Standard deviation and variance. The proposed methodology presented in this paper uses standard deviation and RMSE.

Variance is a factor contributing to the standard deviation. ARIMA based variants for the cumulative cases in some European countries using root mean percentage error, skewness and kurtosis and analyzed. From this methodology, the evaluation metrics of skewness and kurtosis are taken for the proposed methodology which tells the actual trend of the methodology. Susceptible-Exposed-Infectious-Removed (SEIR) model was proposed to detect the trend of Covid 19 in various cities in Mainland China. However, the metrics to evaluate are not clear enough to give a quantitative observation.

Proposed Methodology

An Auto Recurrent Linear Regression (ARLR) Algorithm is proposed based on the skewness of data. The skewness of data observed in the regression defines the slope of the predicted curve. The data obtained is processed using the correlated subspaces of data (John Joseph et al, 2011). The correlation of actual active cases and the cumulative cases show a positive trend. So the number of active cases in time series forecast are taken for regression based forecasting. The population of the country where the data is taken from is approximately 70 million. So the variable nx is set to 70. The number of cases is cumulatively calculated for the period of p days as below.

Cumulative cases $C_n = \sum_{i=1}^{n} c(i)$ (1)

Each day reports r cases recovering and d deaths. So, the cumulative active cases of a day d is calculated as follows.

Cumulative active cases $AC_n = \sum_{i=1}^{n} c(i) - rc(i) - dc(i)$ (2)

Where rc and dc are cumulative recovery and deaths on day i.

$\sum_{i=1}^{n} ACi$ is taken as the series to predict the total active and predicted cases. This series is fed as input to the Auto Recurrent Linear Regression Algorithm. This algorithm is used for predicting active cases and rate of infection. The pseudocode of the ARLR is given below.

Auto Recurrent Linear Regression Algorithm

```
Input: number of millions in population nx, cumulative active cases and transmission rate
Step 1: Input values of cumulative active cases from equation (2)
Step 2: Observe from a window of 5 consecutive values and calculate skewness
Step 3: If skewness is negative, slope is negative and vice versa
Step 4: Perform linear regression with the updated skewness factor $y=mx+c$ where $m$ = skewness x slope. The intercept is subjected to ReLU with $c=max(0,c)$
Step 5: Repeat again from step 2 until the number of active cases go lesser than nx
```

The step 4 in the above algorithm is the modified formula of the linear regression. It includes a calculation of Rectified Linear Unit (ReLU) on the intercept calculated from the linear regression which gives a smoothening of curve towards the actual data. This modified version of formula along with the flow of ARLR algorithm is used to improve the performance metrics. Normally, linear regression will take input of a subseries and will create a single dimensional plane which gives a straight line. While using the proposed algorithm, there are multiple slopes depending on the skewness calculated on the basis of the moving window of five days. When comparing to the traditional linear regression, the proposed algorithm gives a prediction in a trend of skewness observed. The window size of 5 is set after initial experimentation. The initial experimentation gave a convincing progression of data towards the unobserved days.

The infection rate is calculated with three different window configurations. The configurations include window size w = 14, 15 and 21. This window size is inspired from the quarantine interval set by various countries for travelers entering their ports. This window period is actually defined as the incubation period of the pandemic virus to start showing symptoms from a potential victim. For example, if the window period is 14, if one person gets affected on day one and identified by medical tests, 3 people are identified on the day 14, then the 3 people infected on day 14 is considered to be infected by the person on day 1. So the rate of infection is given using the equation below.

$$\text{Rate of infection on day j} = \frac{Total\ infections\ identified\ on\ day\ i}{Total\ infections\ identified\ on\ day\ i-j} \qquad (3)$$

The calculations from the rate of infections on day j = 14, 15 and 21 are calculated and tabulated in the database. This series of rate of infection is also subjected to the proposed ARLR algorithm and the predicted series is obtained. This data predicted is compared with the actual data until the day predicted less than one in millions of populations. The performance of these configurations were identical irrespective of window frame. So j=14 is taken as a configuration to compare with the existing methodologies. The observation of the j=14 gave a sharp peak when compared to j=15 and j=21. So the sharpest peaked configuration is chosen for performance evaluation. The detailed process of experimentation is provided in the next section.

Experimentation

Data regarding the status of covid 19 pandemic was collected for the Kingdom of Thailand. The data schema is given in Figure and collected from the national portal API [25]. This data is stored in a CSV file using Pandas library and calculations are done based on equations 2 and 3. These give rise to four different series of data. From the obtained data, 30% of the observations are used to create initial model of ARLR algorithm. Then the predicted values are recurrently calculated and mapped against actual values observed.

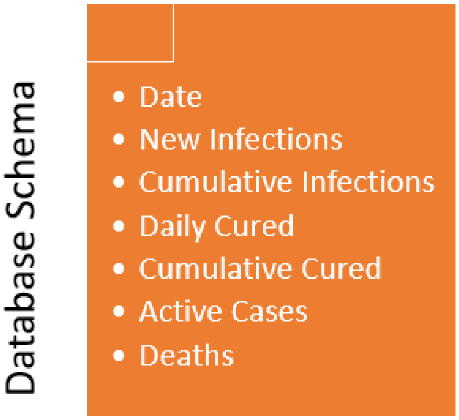

Figure 3: Database schema

This data is then fed as input to the proposed framework involving the recursive exponential regression algorithm. This is done using stochastic simulation of the mathematic formulations in the algorithm. The results are obtained for the expected day of zero active cases and the forecast of null rate of transmission are obtained. The database consists of 100 days of covid 19 parameters as mentioned in the database schema provided in Figure. The experimentation was done by parsing the JSON objects in the given API on an automatic scheduler. This was performed using auto scheduler every day during the pandemic. The proposed auto recursive regression based algorithm was applied on the data collected using the inbuilt mathematical functions in python libraries. The same data was subjected to exponential smoothing, autoregressive moving average and normal linear regression on the cumulative active cases obtained. The performance of the proposed methodology on the data scrapped are illustrated and discussed in the next section. The existing methodologies to compare include linear regression and three factor exponential smoothing. Other methodologies discussed in the literature review are not chosen to evaluate because, most of them are for seasonal trend. Time series forecasting is predominantly done for seasonal trend based data in time series.

The performance metrics taken to evaluate the proposed ARLR algorithm are standard deviation, root mean squared error, skewness and kurtosis. These metrics are done for both active cases and the rate of transmission with window j = 14. These metrics are chosen after studying about the performance metrics of various time series forecasting methodologies.

Standard deviation between the series of predicted against actual data shows the difference between data points between both series compared. Standard deviation of the given series is calculated using the formula given below.

$$\text{Standard deviation of the series SD} = \sqrt{\frac{\sum(x_i-\mu)^2}{N}} \qquad (4)$$

Where N is he total number of data points in the series, x_i is the value of the observed individual data point and μ is the mean of all the data points.

Root Mean Squared Error (RMSE) is calculated as an evaluation metric to find the mean of error magnitudes between the actual and predicted data. RMSE is calculated using the formula given below.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i-A_i)^2}{N}} \qquad (5)$$

Pi is the predicted value and Ai is the actual value of a data point in a series with N data points.

Skewness is a parameter calculated as a part of proposed ARLR algorithm but used as a performance metric to measure the overall trend of the given prediction algorithm as it gives the extent of symmetric graph. It is calculated using the formula given below.

$$\text{Skewness} = \frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{s}\right)^3 \qquad (6)$$

Kurtosis is finally measured to define whether the series is leptokurtic or platykurtic as it is the fourth standardized moment. It is calculated using the formula as follows.

$$\text{Kurtosis} = \text{E}\left[\left(\frac{x-\mu}{\sigma}\right)^4\right] \qquad (7)$$

Results and Discussion

The data collected over the pandemic period was analyzed and the actual cases on daily basis is visualized in the figure 4 and the rates of transmission are given in figure 5.

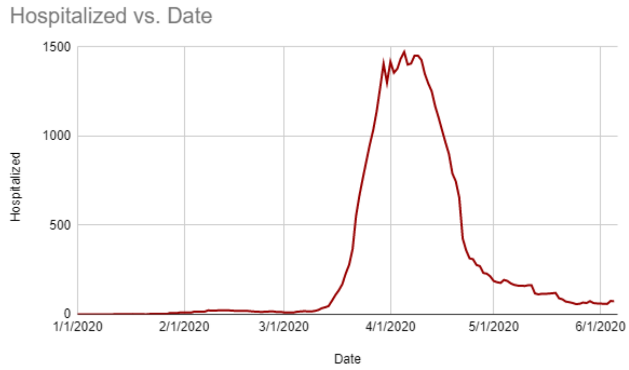

Figure 4: Actual cases observed

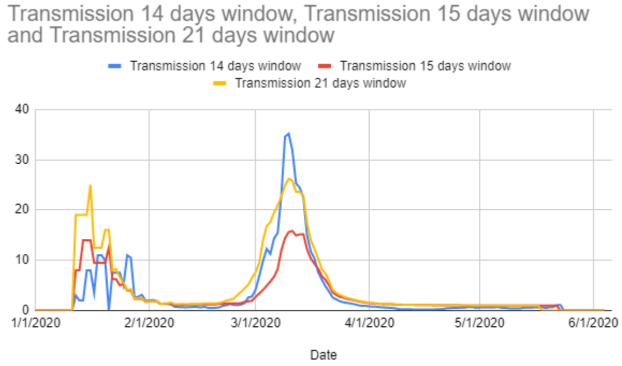

Figure 5: Transmission rate based on the active, recovered cases and deaths

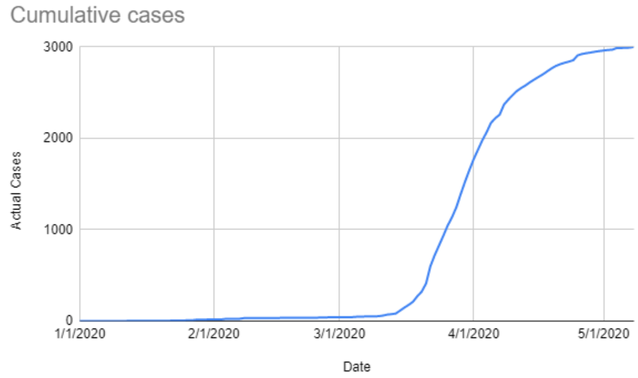

Figure 6: Cumulative covid 19 cases in Thailand

The proposed methodology is evaluated using the metrics of standard deviation and root mean square error (RMSE). Standard deviation and RMSE are mostly used to validate the time series forecasting between actual and predicted values. The quantitative results of the actual vs predicted active cases is provided in figure 7 and the actual vs predicted rate of transmission is given in figure 8. The data was analyzed until the day when actual active cases came down to less than one in a million population. The data collected does not include those detected from the state quarantine facilities which hosts those returning from other countries after lock down of air transport.

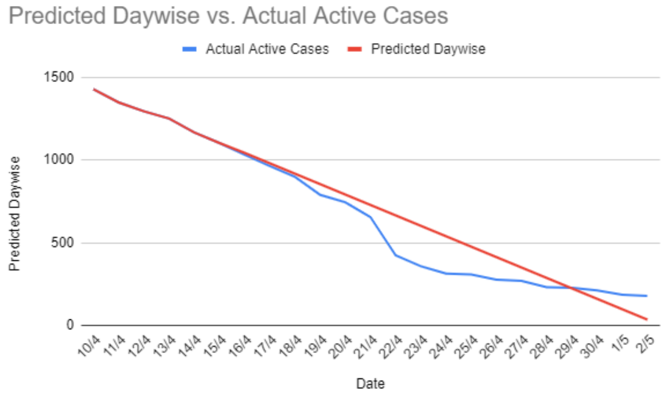

Figure 7: Actual active cases against predicted active cases

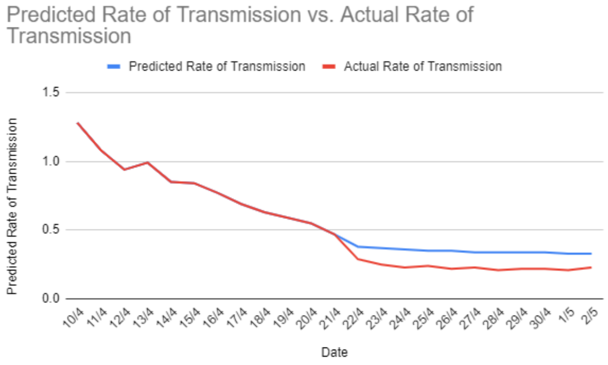

Figure 8: Actual rate of transmission against predicted rate of transmission

As mentioned in the section of proposed methodology, the line trend of predicted rate of transmission is not a straight line while similar observation is also obtained in active cases. This is because of the recurrent model creation using the ARLR algorithm and the effectiveness are evident from the results. The results of the evaluation metrics on the proposed and existing methodologies are given in table 1.

Parameter	**Proposed Vs Actual**	**Actual vs ARIMA (Benvenuto et al)**	**Actual vs (Yang et al)**
Standard Deviation of active cases	**430.22**	464.22	498.54
Standard deviation of rate of transmission	**0.31**	0.55	0.68
RMSE of actual cases	**27 ± 0.01**	34 ± 28	49 ± 59
RMSE of rate of transmission	**0 ± 0.77**	0 ± 0.88	0 ± 0.91
Skewness of active cases	**0.00134**	0.000124	0.00102
Skewness of rate of infection	**0.927**	0.734	0.697
Kurtosis of active cases	**-1.522**	-1.1003	0.0063
Kurtosis of rate of infection	**-0.211**	-0.0159	0.0045

Table 1: Performance evaluation of proposed methodology against normal regression and stationary forecast methods

It is clearly evident that the auto recurrent linear regression holds better than the existing non seasonal methods like linear regression and exponential smoothing. The standard deviation of predicted active cases is lesser than those of the existing methods against the actual data.

A positive skew is observed with a higher magnitude is observed when it comes to the predicted actual cases and rate of infection. Platykurtic trend is observed in proposed methodology, while a mix of platykurtic and leptokurtic trends are observed in all the existing methodologies. The leptokurtic trend on exponential smoothing proves that the time series forecasting for pandemic cannot be seen seasonal.

The performance of proposed methodology is not compared with many of the time series forecast analysis because, the trend of pandemic is not seasonal and stationary. This has an exponential growth based on the previous day's data.

Conclusion

In this paper, an auto recurrent linear regression algorithm for non-seasonal trends of time series forecasting is proposed. This proposed methodology is developed after comparing with other methodologies in non-seasonal trend of time series data. This methodology is applied to a real-time covid dashboard data and evaluated to check the prediction of the pandemic progression. The progression of pandemic is evaluated for the accumulation and dissimilation of active cases over the period of the pandemic. Secondly, the transmission rate of infection is also checked using this technique until the active cases go equal to or less than one per million populations. Normal linear regression and three factor exponential smoothing are compared with the proposed methodology. These methods are chosen due to their non-seasonal forecasting trends. The proposed methodology is found to perform better than the existing methodologies. This claim of the proposed methodology is justified using the performance metrics of Standard Deviation, Root Mean Square Error, Skewness and Kurtosis.

References

T. P. Velavan and C. G. Meyer, 2020, “The COVID-19 epidemic,” Trop. Med. Int. Heal., vol. 25, no. 3, p. 278.

D. J. Bartholomew, 1971, “Time series analysis forecasting and control,” J. Oper. Res. Soc., vol. 22, no. 2, pp. 199–201.

P. J. Brockwell and R. A. Davis, 1996, Introduction to time series and forecasting. springer.

H. Akaike, 1974, “A new look at the statistical model identification,” IEEE Trans. Automat. Contr., vol. 19, no. 6, pp. 716–723.

N. A. Gershenfeld and A. S. Weigend, 1994, Time Series Prediction: Forecasting the Future and Understanding the Past: Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis Held in Santa Fe, New Mexico, May 14-17, 1992. Addison-Wesley.

F. Takens, 1981, “Detecting strange attractors in turbulence,” in Dynamical systems and turbulence, Warwick 1980, Springer, pp. 366–381.

N. I. Sapankevych and R. Sankar, 2009, “Time series prediction using support vector machines: a survey,” IEEE Comput. Intell. Mag., vol. 4, no. 2, pp. 24–38.

M. Ghassemi et al., 2015, “A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data,” in Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 446–453.

H. J. Kam, J. O. Sung, and R. W. Park, 2010, “Prediction of Daily ED Patient Numbers,” Healthc. Inform. Res., vol. 16, no. 3, pp. 158–165.

Z. Hu, Q. Ge, L. Jin, and M. Xiong, 2020, “Artificial intelligence forecasting of covid-19 in china,” arXiv Prepr. arXiv2002.07112.

D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi, 2020 “Application of the ARIMA model on the COVID-2019 epidemic dataset,” Data Br., vol. 29, p. 105340.

S. Deb and M. Majumdar, “A time series method to analyze incidence pattern and estimate reproduction number of COVID-19,” arXiv Prepr. arXiv2003.10655, 2020.

K. Mizumoto and G. Chowell, 2020, “Transmission potential of the novel coronavirus (COVID-19) onboard the diamond Princess Cruises Ship, 2020,” Infect. Dis. Model., vol. 5, pp. 264–270.

F. Petropoulos and S. Makridakis, 2020, “Forecasting the novel coronavirus COVID-19,” PLoS One, vol. 15, no. 3, p. e0231236.

Z. Yang et al., 2020, “Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions,” J. Thorac. Dis., vol. 12, no. 3, p. 165.

F. J. John Joseph, R. T, and J. J. C, 2011 “Classification of correlated subspaces using HoVer representation of Census Data,” International Conference on Emerging Trends in Electrical and Computer Technology, pp. 906–911.

M. of P. Health, “Department of Disease Control,” th-stat.com, 2020. https://covid19.th-stat.com/th/api.