Time-series Forecast of the COVID-19 Pandemic Using Auto Recurrent Linear Regression

Ferdin Joe John Joseph

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

Corresponding Author: ferdin@tni.ac.th

Submitted : 31-08-2021

Revised : 01-01-2022

Accepted : 21-03-2022

ABSTRACT

Since the beginning of 2020, the COVID-19 pandemic has severely affected the economy and lifestyle of people. Various data analytics have been performed on data obtained from various sources. These analytics include symptom prediction, time-series forecasting, and impact analyses. Forecasting the end of the pandemic remains a challenge for many countries. Time-series forecasting models have been proposed for various applications. However, a nonseasonal and nonstationary forecasting method is required for predicting the progression of the pandemic. An auto regressive linear regression algorithm was proposed using the COVID-19 data of a certain geography. The results of the proposed methodology are convincing when compared with the nonseasonal and nonstationary existing methodologies, such as linear regression and exponential smoothing variants. The standard deviation and root mean square error of the proposed method were 430.22 and 0.31, respectively, for active cases, and 27.01 and 0.77 for the rate of transmission with positive skew and platykurtic trend, respectively.

Keywords: Covid 19; Time Series Forecasting; Recurrent Linear Regression; Pandemic Forecasting; Auto Recurrent Linear Regression.

INTRODUCTION

The COVID-19 pandemic drastically paralyzed the economy, health, and lifestyle of people in the first half of 2020. The end of the pandemic and return to normalcy was highly anticipated. The COVID-19 pandemic was caused by a variant of the severe acute respiratory syndrome (SARS) that originated in China in 2019 (Velavan and Meyer, 2020). Many mathematical models have been proposed for predicting future trends in pandemics. The most critical mathematical basis is time-series forecast analysis. Typically, time-series forecast analysis is performed for seasonal, stationary, and exponential data. Pandemic data form a different trend from seasonal data, as displayed in the bell diagram in Figure 1.

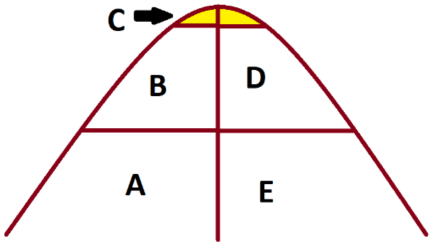

Figure 1: Active cases curve during a pandemic

Parts A and B have a positive slope, whereas parts D and E have a negative slope, which denotes the rise and fall of active cases, respectively, during the pandemic. Hump C has a near-zero slope, which marked the peak impact caused by the pandemic.

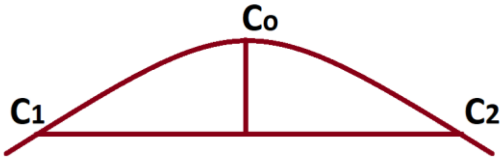

Figure 2: Hump C of the pandemic curve

Tangent β_1 corresponding to arc C_1C_o indicates the near peak, and tangent β_2 corresponding to arc C_oC_2 denotes the identifier, which leads to a decrease in the impact of the pandemic. A pandemic curve in a particular geography can have single or multiple humps, and each hump includes parts mentioned in Figures 1 and 2. These humps play a crucial role in predicting the trend of pandemics using time-series forecasting. The cumulative cases over the pandemic curve follow an exponential trend, but this trend is not helpful in predicting the end of the pandemic.

The proposed methodology is explained in the Experimentation section, the results of the experiments are illustrated in the Results and Discussion section, and finally, the conclusion is presented.

Related Work

Time-series analysis and forecasting have been performed for decades using a large corpus of mathematical formulations. A univariate discrete time series is a stochastic process with probabilistic terms. This method is an autoregressive model that deals with previous values in the series, known as the autoregressive integrated moving average (ARIMA). Brockwell and Davis (1996) used an autocorrelation function to manage the municipal solid waste forecasting in Spain. The peaks observed using this method revealed 12 lag units that perfectly matched the seasonal trend. For a nonseasonal trend and stationary behavior, Takens (1981) proposed a stationary method of autoregressive moving average (ARMA). In some nonlinear time-series analysis methods (Gershenfeld and Weigend, 1994), discrete maps are used. These methods are simple and proposed in the genesis of time-series analysis, which is used in various applications. The scope of the time series is discussed extensively [6], and the primary applications are stock prediction, electricity utility, and business applications.

1. Medical data forecasting

The severity of illness was predicted from ICU data using a multivariate time-series modeling approach (Sapankeiwich and Sankar, 2009) integrating the Gaussian process. The results were evaluated using 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, Kam et al. (2010) used autoregressive and multinomial distributions of the number of calls to predict the incoming number of calls to the emergency service. Because emergency health data are seasonal, the methodology for pandemic progression should be refined to go along with a backpropagation algorithm.

2. COVID-19 time-series analysis

Few studies have focused on the COVID-19 pandemic in correlation with medical data-based applications. Hu et al. (2020) developed a stacked encoder for the transmission of epidemics using the data available in China during the receding phase. The average error was calculated to verify the effectiveness of the stacked encoder. The ARIMA model was used on cloud data (Benvenuto et al., 2020), and a correlellogram was presented on the proposed method. The rate of infection was statistically analyzed. Deb and Majumdar (2020) used the time-dependent pattern and evaluated for the root mean square error (RMSE) between actual and predicted data. However, this domain-specific time series did not follow a seasonal trend. A split-based analytics approach was proposed by Mizumoto and Chowell (2020), who presented the pandemic progression between three groups of people using an estimate of the mean reproduction number. The psychological aspects of the pandemic progression (Petropoulos and Makridakis, 2020) were presented with a timeline of events and their impact on the people. This finding was similar to that of a quantitative study conducted on people. A deep learning LSTM method was proposed to predict the time series of pandemics in Canada (Yang et al., 2020), and the performance of the method was evaluated using RMSE. Most methodologies were evaluated using the RMSE, standard deviation, and variance. In the proposed methodology, the standard deviation and RMSE were used. Variance was a factor contributing to the standard deviation. ARIMA-based variants for the cumulative cases in some European countries were analyzed using root mean percentage error, skewness, and kurtosis. From this methodology, the evaluation metrics of skewness and kurtosis were obtained for the proposed methodology. This phenomenon indicates the actual trend of the methodology. The susceptible-exposed-infectious-removed (SEIR) model was proposed to detect COVID-19 in various cities in Mainland China. However, the metrics to be evaluated were not sufficiently clear to provide a quantitative observation.

Proposed Methodology

An auto-recurrent linear regression (ARLR) algorithm was proposed based on the skewness of data. The skewness of the data observed in the regression analysis defines the slope of the predicted curve. The data obtained were processed using the correlated subspaces of data (John Joseph et al., 2011). The correlation between the active cases and cumulative cases revealed a positive trend. Therefore, the number of active cases in the time-series forecast was used for regression-based forecasting. The population of the country from which the data were obtained is approximately 70 million. Therefore, n_x was set to 70. The number of cases was cumulatively calculated for the period of p days as follows:

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

Each day, r cases of recovery and death are reported. Thus, the cumulative active cases of day d are calculated as follows:

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

where r_c and d_c are cumulative recovery and deaths on day i.

Here, $\sum_{i=1}^{n} ACi$ is the series used to predict the total active and predicted cases. This series is inputted to the ARLR algorithm. This algorithm was used to predict the active cases and rate of infection. The pseudocode of ARLR is as follows:

ARLR Algorithm

```
Input: Number of millions in population $n_x$, cumulative active cases and transmission rate
Step 1: Input values of cumulative active cases from equation (2)
Step 2: Observe from a window of five consecutive values and calculate skewness
Step 3: If skewness is negative, the 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 $n_x$
```

Step 4 in the above algorithm is a modified formula of linear regression. This step includes the calculation of the rectified linear unit (ReLU) on the intercept calculated from linear regression, which smoothens the curve toward the actual data. This modified version of the formula, along with the flow of the ARLR algorithm, was used to improve performance metrics. Normally, linear regression takes the input of a subseries and creates a single-dimensional plane that provides a straight line. When using the proposed algorithm, multiple slopes exist depending on the skewness calculated based on a moving window of 5 days. When compared with conventional linear regression, the proposed algorithm predicts the observed skewness trend. A window size of 5 was set after the initial experimentation. The initial experimentation provided a convincing progression of data toward unobserved days.

The infection rate was calculated using three window configurations. The configurations included window sizes of w = 14, 15, and 21. This window size is inspired by the quarantine intervals set by various countries for travelers to enter their ports. This window period is defined as the incubation period of the pandemic virus to begin showing symptoms from a potential victim. For example, if the window period is 14, if one person is affected on day 1 and identified by medical tests, 3 people are identified on day 14, and then three people infected on day 14 are considered to be infected by the person on day 1. Therefore, the infection rate is given by the following equation:

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

The rate of infections on days j = 14, 15, and 21 were calculated and tabulated in the database. This series of infection rates was subjected to the proposed ARLR algorithm, and the predicted series was obtained. The predicted data were compared with the actual data until the day less than one was predicted in millions of populations. The performances of these configurations were identical irrespective of the window frame. Therefore, *j* = 14 is considered as a configuration for comparison with existing methodologies. The observation of *j* = 14 revealed a sharp peak compared with *j* = 15 and *j* = 21. Therefore, the sharpest peak configuration was selected for performance evaluation. The experimental process is described in detail in the next section.

Experimentation

Data on the status of the COVID-19 pandemic were collected from the Kingdom of Thailand. The data schema is displayed in Figure and was collected from the national portal API [25]. These data are stored in a CSV file using the Pandas library, and calculations are performed based on Equations 2 and 3. This phenomenon details four data series: from the data obtained, 30% of the observations were used to create the initial model of the ARLR algorithm. The predicted values were then recurrently calculated and mapped against the actual observed values.

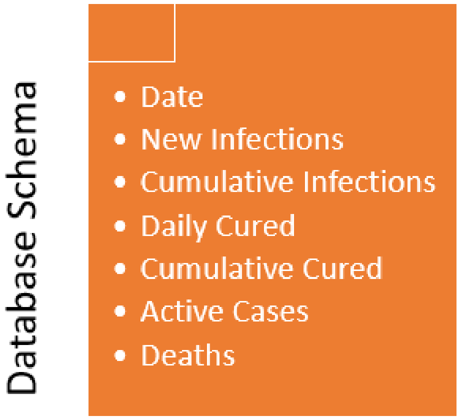

Figure 3: Database schema

These data are then inputted to the proposed framework, which involves a recursive exponential regression algorithm. This process is performed using a stochastic simulation of mathematical formulations in the algorithm. The results were obtained for the expected day of zero active cases, and the forecast of the null rate of transmission was obtained. The database consists of 100 days of COVID-19 parameters as mentioned in the database schema provided in Figure. The experiment was performed by parsing the JSON objects in the given API on an automatic scheduler. This process was performed using an auto-scheduler daily during the pandemic. The proposed auto-recursive regression-based algorithm was applied to the data collected using built-in mathematical functions in Python libraries. The same data were subjected to exponential smoothing, autoregressive moving average, and normal linear regression for cumulative active cases obtained. The performance of the proposed methodology on scrapped data is illustrated and discussed in the next section. The existing methodologies for comparison include linear regression and three-factor exponential smoothing. Other methodologies discussed in the literature review were not selected for evaluation because most reviews were used for seasonal trends. Time-series forecasting is predominantly performed for seasonal trend-based data in time series.

The performance metrics used to evaluate the proposed ARLR algorithm were the standard deviation, RMSE, skewness, and kurtosis. These metrics were performed for both active cases, and the rate of transmission with window *j* = 14. These metrics were selected after studying the performance metrics of various time-series forecasting methodologies.

The standard deviation between the series of predicted versus actual data revealed the difference between the data points of the two series compared. The standard deviation of the given series was calculated using the following formula:

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

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

The RMSE was calculated as an evaluation metric to determine the mean of the error magnitudes between actual and predicted data. The RMSE was calculated using the following formula:

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

Here, P_i is the predicted value and A_i is the actual value of a data point in a series with N data points.

Skewness is a parameter that is calculated as a part of the proposed ARLR algorithm but is used as a performance metric to measure the overall trend of the given prediction algorithm because it provides the extent of the symmetric graph. The parameter was calculated using the following formula:

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

Kurtosis is measured to define whether the series is leptokurtic or platykurtic, as it is the fourth standardized moment. Kurtosis was calculated using the following formula:

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

Results and Discussion

The data collected over the pandemic period were analyzed. The actual cases on a daily basis are displayed in Figure 4, and the rates of transmission are displayed in Figure 5.

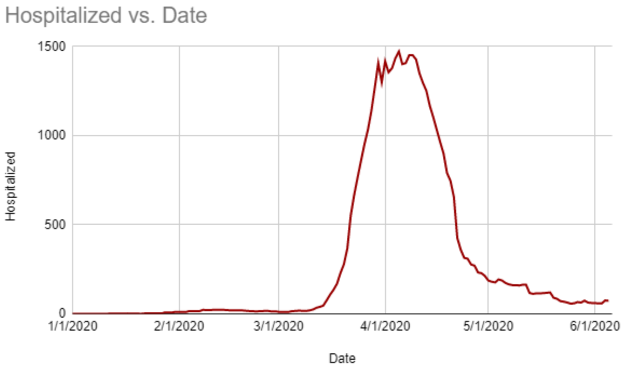

Figure 4: Actual cases observed

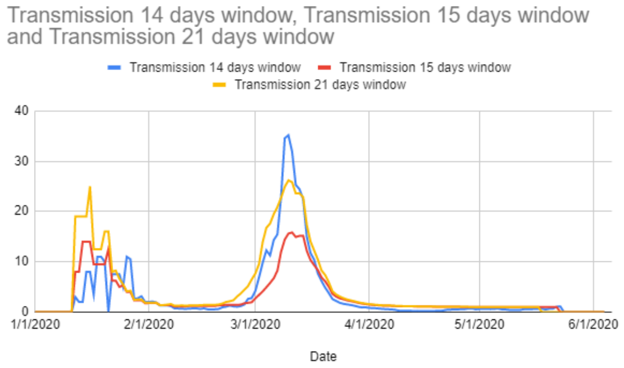

Figure 5: Transmission rate based on the active, recovered cases, and deaths

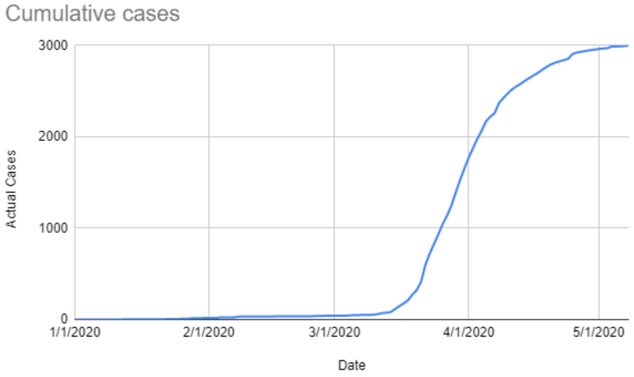

Figure 6: Cumulative COVID-19 cases in Thailand

The proposed methodology was evaluated using the standard deviation and RMSE metrics. The standard deviation and RMSE were used to validate time-series forecasting between actual and predicted values. The quantitative results of the actual versus predicted active cases are displayed in Figure 7, and the actual versus predicted rate of transmission is displayed in Figure 8. The data were analyzed until the day when actual active cases decreased to less than one million people. The data collected did not include cases detected from the state quarantine facilities that host people returning from other countries after the lockdown of air transport.

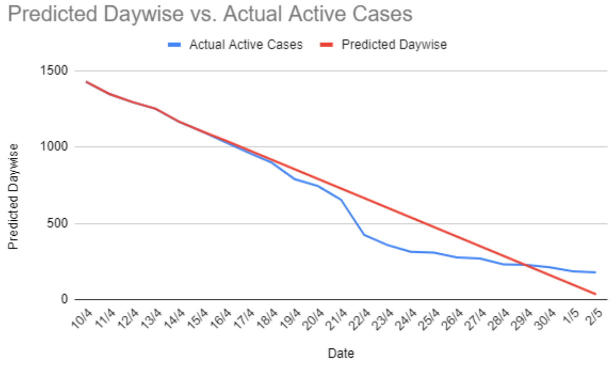

Figure 7: Actual active cases against predicted active cases

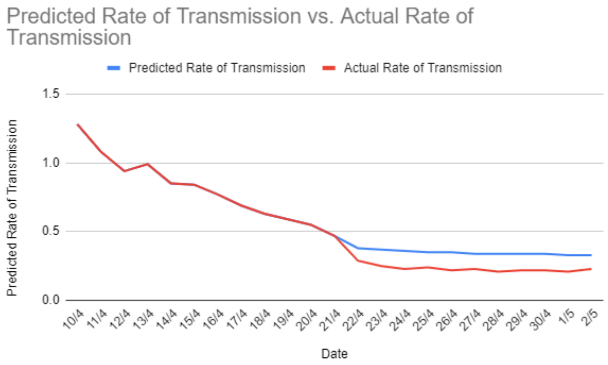

Figure 8: Actual rate of transmission against the predicted rate of transmission

As mentioned in the section on the proposed methodology, the line trend of the predicted rate of transmission is not a straight line, and similar observations are obtained in active cases. This phenomenon can be attributed the recurrent model creation using the ARLR algorithm, and its effectiveness is evident from the results. The results of evaluation metrics for proposed and existing methodologies are listed 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 the rate of transmission	**0 ± 0.77**	0 ± 0.88	0 ± 0.91
Skewness of active cases	**0.00134**	0.000124	0.00102
Skewness of the rate of infection	**0.927**	0.734	0.697
Kurtosis of active cases	**-1.522**	-1.1003	0.0063
Kurtosis of the rate of infection	**-0.211**	-0.0159	0.0045

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

The auto-recurrent linear regression is superior than existing nonseasonal methods such as linear regression and exponential smoothing. The standard deviation of the predicted active cases against the actual data was less than that of existing methods.

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

The performance of the proposed methodology was not compared with that of many time-series forecast analyses because the pandemic trend was not seasonal or stationary. This phenomenon has exponential growth based on data from the previous day.

CONCLUSION

In this study, an ARLR algorithm was proposed for nonseasonal trends in time-series forecasting. This proposed methodology was developed after comparison with other methodologies in the nonseasonal trend of time-series data. This methodology was applied to real-time COVID-19 dashboard data to evaluate the prediction of pandemic progression. The progression of pandemic was evaluated for the accumulation and dissimilation of active cases over the period of the pandemic. Second, the transmission rate of infection was checked using this technique until the active cases were equal to or less than one per million population. Normal linear regression and three-factor exponential smoothing were compared with the proposed methodology. These methods were selected because of their nonseasonal forecasting trends. The proposed methodology outperformed existing methodologies. The performance of the proposed methodology was validated using standard deviation, RMSE, 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, article no. 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, article no. 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, T. Ravi, and C. J. Justus, 2011 “Classification of correlated subspaces using HoVer representation of Census Data,” in International Conference on Emerging Trends in Electrical and Computer Technology, Nagercoil, India, pp. 906–911.

Ministry of Public Health, Department of Disease Control, th-stat.com, 2020. https://covid19.th-stat.com/th/api.