Received: 12 January 2024 *Accepted: 09 June 2024*

Assessing Statistical Methods for Generating Forecasts for COVID-19

[1]Siriwardena, S. M. D. G. A, [1]Dharmaratne, G.[1], & [2]Amaratunga, D.

gethsiriwardena@gmail.com, sameera@stat.cmb.ac.lk, damaratung@yahoo.com
[1]Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka.
[2]Princeton Data Analytics, New Jersey, USA.

Abstract

The COVID-19 pandemic, a persistent global health emergency that has affected almost all facets of daily life, was initially discovered in Wuhan, China, in December 2019. Since that time, the virus has rapidly spread over the globe, causing serious social and economic upheavals necessitating the need for reliable forecasting methods. This study compares ten distinct models to predict the number of confirmed COVID-19 cases in Sri Lanka, aiming to assess the performance of statistical models using limited and volatile real-world data characterized by trends, random peaks, and autocorrelations. In addition to the classical ARIMA model, various smoothing and filtering techniques were explored to capture the unique characteristics of the data. The model consistencies in multiple-day predictions were demonstrated, and robust evaluation criteria, along with non-robust measures, were utilized to enhance the effectiveness of the evaluation process. The results highlight the effectiveness of traditional smoothing and filtering strategies such as Simple Exponential Smoothing, Holt's Exponential Smoothing, and the Smoothing Splines technique coupled with the ARIMA model. This study also discovered that the ARIMA model, when applied directly to the original data without using any smoothing or filtering approaches, failed to forecast adequately, thereby demonstrating the insufficiency of the ARIMA model on its own to provide credible forecasts when given a volatile set of data.

Keywords: Arima, Smoothing, Time series, Trend analysis.

Introduction

The COVID-19 pandemic has had a profound and lasting impact on daily life worldwide, including Sri Lanka. Given the need to implement preventive measures to mitigate the impact of the pandemic, a reliable forecasting approach becomes crucial. It helps public health experts allocate resources, prepare for outbreaks, establish healthcare systems, and make decisions on lockdowns, and other mitigation plans. However, available data, such as the relatively short time series data set that is used in the current study,are essentially dominated by significant autocorrelations, trends, and outliers, rather than aspects like seasonal variations. Therefore, standard time series methodologies have some challenges

in performing well. This study, thus, aims to compare the forecasting performance of alternative statistical models for predicting COVID-19 cases in Sri Lanka, offering insights into optimal methods for diverse circumstances.

In general, COVID-19 data are characterized by dominant trends. Smoothing and filtering techniques can effectively address these issues in time series data. For trend capture, LOWESS (Cleveland, 1979) and spline (Craven & Wahba, 1979) models have been widely used. LOWESS smoothing has been applied to represent the trend in a study of tetanus vaccines and TBE, (Hainz et al., 2005), while cubic spline smoothing based on a stochastic state space model has shown promising results in predicting COVID-19 cases and fatalities (Gecili et al., 2021). Spline smoothing techniques have also been used in conjunction with machine learning methods for effective forecasting of COVID-19 cases. Exponential smoothing methods, such as Holt's model, have also been commonly employed and found to be reliable for short-term forecasts of daily COVID-19 cases (Martinez et al., 2020). Furthermore, time series data characterized by zero-inflation (Tawiah et al., 2021) were analysed in modelling of COVID-19 deaths in Ghana.

However, there is a need for further research on combining ARIMA with other filters or smoothers for improved predictions, as well as a scarcity of literature specifically focusing on COVID-19 cases and fatalities in Sri Lanka.

Materials and Methods

The dataset used in this study was sourced from the website https://ourworldindata.org/. The variable, namely 'new cases', was utilized to compare different models for predicting COVID-19 cases in the context of Sri Lanka. The data collection period began on January 27, 2020, and is updated daily. The study covers a time span of over two years until the decline of COVID-19 cases in Sri Lanka, comprising a total of 892 data points until July 6, 2022. The time series considered in this study is illustrated in Figure 1 and Figure 2.

Figure 1.
Plot of COVID-19 Cases.

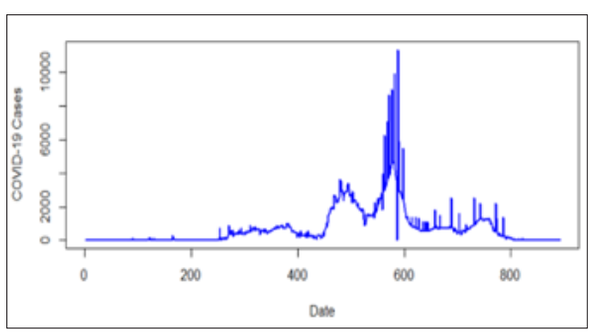
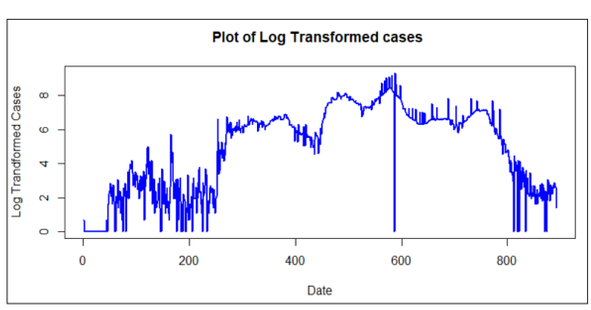


Figure 2.
Plot of Log Transformed COVID-19 Cases.



The COVID-19 trajectory for cases exhibits an increasing trend from low levels, followed by a decline towards zero after reaching a peak. Moreover, there are no strong seasonal patterns observed in the plot or in the ACF analysis. However, fragmentary trends and significant

autocorrelations at higher lags suggest the presence of short-term correlations and non-stationarity. Notably, there are prominent outliers, emphasizing the importance of robust evaluation metrics. Considering these characteristics, trend components and sporadic peaks dominate the series, potentially posing challenges for statistical modelling. Nevertheless, the application of smoothing and filtering techniques may help address these complexities. The time series plot exhibits significant variability, outliers, and extended tails. To mitigate these issues, a log transformation was applied, effectively reducing skewness and variability in the data. However, before comparison between models, all the models were converted back to the original scale. Since there were days with no cases (65 days), when the log transformation was applied, an extra 1 was added to all the cases in each day.

To achieve the objective of comparing methods that can capture the properties of the data and to evaluate their forecasts, in this study, the training data set is continuously updated with each realization of the testing data, ensuring adaptiveness and continuity throughout the analysis (Table 1).

Model implementation

The following techniques were utilized in this study for comparison purposes,

01. **ARIMA on the original data (RAW)**
First, the ARIMA model was fitted on the raw data as a baseline model to gauge how different (or better) the other models performed against it. The auto. arima() function in R was used to fit the ARIMA model.

02. **ARIMA on the log transformed data (LOG)**
The log transformation was applied to all other models in this analysis, excluding the previous case and ARIMA model was also fitted to this.

03. **ARIMA on the log transformed, LOWESS (linear fit) smoothed data (LOWESS)**
The LOWESS smoother was applied to the logged data set using the 'lowess()' function in R. A two-week moving window was selected as the smoother's bandwidth to strike a balance between smoothness and agreement with local characteristics.

Table 1.
Process of forecasting.

	Range of fitting	Predicting	
Next day predictions	1:x	x+1	x;600,601,602,…,891
One-week ahead predictions	1:x	x+7	x;600,601,602,…,885
One-month (*4-weeks*) ahead predictions	1:x	x+28	x;600,601,602,…,864

04. **ARIMA on the log transformed, LOWESS (quadratic fit) smoothed data (LOESS**)
The 'loess()' function in R was utilized to implement this approach. Similar to LOWESS, a two-week window was used for smoothing. However, unlike LOWESS, a quadratic polynomial was fitted instead of linear. To mitigate the impact of outliers on the estimation process, a re-descending M estimator with Tukey's bi-weight function was employed.

05. **ARIMA on the log transformed, SPLINE smoothed data (SPLINE)**
In this study, a cubic smoothing spline was utilized for spline smoothing using the 'smooth.spline()' function in R. The default smoothing parameter, which captures global characteristics, was adjusted to strike a balance between adhering to the time series and achieving suitable smoothness. For cases, the best degrees of freedom were determined as 40.

06. **ARIMA on log transformed, Running Means data (RM)**
A 7-day moving average trend was calculated for the log-transformed data using the 'filter()' function in R.

07. **Simple Exponential Smoothing on log transformed data (SES)**
Simple exponential smoothing was performed using the 'ses()' function from R. The estimated smoothing values for level ('alpha') were 0.34 for cases. The smoothing parameters and initial values were optimized together. The 'initial' argument was set to 'optimal' for selecting initial state values. Forecasting periods ('h') of 1, 7, and 28 were chosen to match the desired time frames.

08. **Holt's Exponential Smoothing (without damping the trend) on log transformed data (HOLT)**
The 'forecast' package was used to implement the 'holt()' function .The smoothing parameters (α for level, β for trend) were automatically estimated: α was around 0.34 while β was around 0.0004.

09. **Holt's Exponential Smoothing (damping the trend) on log transformed data (HOLT (D))**
This second model, HOLT(D), is a derivation of the previous method incorporating a damped trend ('damped' set to 'True'). The estimated damping parameter (φ) was approximately 0.80.

10. **ARIMA on log transformed, Savitzky-Golay Filtered data (SG)**
This was performed in R using the 'sgolayfilt()' function from the 'filter. sgolay' package. The two most critical parameters that must be predetermined are the half-width and polynomial degree in this function. If the half width is high, the smoothness will usually be excellent at the expense of flattening the sharp peaks. A cubic polynomial with a length of 15 units (half width of 7) was used in this study.

11. **Artificial Neural Networks on the log transformed data (ANN)**
The ANN model is chosen as a

benchmark for comparison with other models. The 'nnetar()' function is used to identify NNAR (Neural Network Auto Regression) models, which are feed-forward neural networks with a single hidden layer and lagged inputs for univariate time series forecasting. The fitted models for non-seasonal data are designated as NNAR(p, k), with NNAR(8, 4) for cases in which 20 networks with random starting values are trained, and their forecasts are averaged.

The models were applied to the testing data, and their forecasting performance was evaluated at various time intervals, while assessing the F using Median Absolute Deviation (MAD), Root Mean Squared Error (RMSE), and Robust Root Mean Squared Error (RRMSE). When assessing results, more weight was given to MAD and RRMSE since they are less influenced by outliers in the prediction period. The best-performing models were identified based on their scores across multiple evaluation metrics and presented accordingly.

Results and Discussion

All 11 methods are compared at different time intervals (see appendix A). This includes one-day, one-week and one-month forecasts with the objective of determining the most effective method for each time frame. The various statistics used to compare the methods under three different scenarios are shown in Table 2.

Table 2.
Forecast evaluation of predictions of COVID-19 cases.

(α) **Next day predictions**

Methods	Statistical indicators		
	RMSE	**MAD**	**RRMSE**
RAW	298.9	103.7	120.2
LOG	265.1	61.8	74.5
LOWESS	243.1	25.5	24.0
LOESS	237.3	16.4	17.2
SPLINE	237.2	48.7	60.5
RM	218.7	45.7	59.3
SG	228.1	66.3	77.9
SES	196.6	42.5	53.5
HOLT	200.1	44.1	63.6
HOLT(D)	196.9	42.4	54.0
ANN	183.9	38.8	47.1

(β) **One week ahead predictions**

Models	Statistical indicators		
	RMSE	**MAD**	**RRMSE**
RAW	306.3	105.0	125.5
LOG	329.3	117.1	139.3
LOWESS	252.5	41.5	42.4
LOESS	277.4	71.4	77.4
SPLINE	238.4	45.7	58.1
RM	289.3	83.3	97.9
SG	315.1	105.4	126.7
SES	257.1	60.9	76.4
HOLT	266.5	73.5	117.9
HOLT(D)	257.0	60.9	76.4
ANN	272.2	60.3	72.8

(c) One month ahead predictions

Models	Statistical indicators		
	RMSE	MAD	RRMSE
RAW	475.7	306.8	337.5
LOG	516.1	247.5	296.8
LOWESS	397.5	246.9	225.6
LOESS	455.4	263.4	264.0
SPLINE	280.8	102.4	111.6
RM	460.8	235.5	270.5
SG	347.2	144.3	165.4
SES	256.6	52.9	59.3
HOLT	328.4	218.5	278.9
HOLT(D)	256.5	52.18	59.1
ANN	297.8	114.5	162.4

Among the models tested, LOESS, LOWESS, smoothing splines, Simple Exponential Smoothing, and the Holt exponential smoothing model with the damped trend were the most reliable for COVID-19 forecasting in Sri Lanka. These models consistently outperformed others across different prediction time frames, especially with regard to the outlier-resistant criteria MAD and RRMSE. For short-term predictions, LOWESS and LOESS models demonstrated exceptional performance. Meanwhile, for mid- to long-term forecasts, the smoothing splines, Simple Exponential Smoothing and Holt exponential smoothing with the damped trend models showed strong predictive capabilities.

Conclusions

This study contributes to the ongoing efforts in predicting the spread of COVID-19 by identifying effective statistical models for forecasting cases in Sri Lanka. The evaluation of various models provides insights into their suitability for different scenarios and future pandemics. The analysis of the brief time series data revealed the presence of trends and random peaks, necessitating the use of techniques such as filtering and smoothing to capture these patterns. Among the models examined, the raw ARIMA model demonstrated the least effectiveness, struggling to handle the complexities and nonlinear patterns present in real-world data. Traditional smoothing techniques, along with ARIMA, outperformed other filtering and deep learning approaches. Notably, exponential smoothing techniques, smoothing splines, LOWESS, and LOESS models exhibited strong performance, sometimes surpassing even complex machine learning models like ANN. This study emphasizes the efficacy of smoothing and filtering approaches over complex machine learning models. Furthermore, it is important to note that under the availability of large volumes of data, this analysis may be further enhanced by incorporating a validation set into the analysis. However, careful consideration of the temporal evolution of autocorrelation is necessary to ensure the credibility of the validation set and its appropriateness for evaluating model performance. Finally, further research incorporating external factors and regular updates is recommended to enhance the accuracy of forecasts.

References

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. In *Source: Journal of the American Statistical Association.* 74 (368). 829-836.

Craven, P., & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematica*, *31*(4), 377–403.

Gecili, E., Ziady, A., & Szczesniak, R. D. (2021). Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. PLOS ONE, 16(1), e0244173.

cMartinez, E. Z., Aragon, D. C., & Nunes, A. A. (2020). Short-term forecasting of daily COVID-19 cases in Brazil by using the holt's model. *Revista Da Sociedade Brasileira de Medicina Tropical*, *53*, 1–5.

Tawiah, K., Iddrisu, W. A., & Asampana Asosega, K. (2021). Zero-inflated time series modelling of covid-19 deaths in Ghana. *Journal of Environmental and Public Health*, 2021, 1–9.

Appendix A

Figure 1.
Fit of the models of COVID-19 cases.

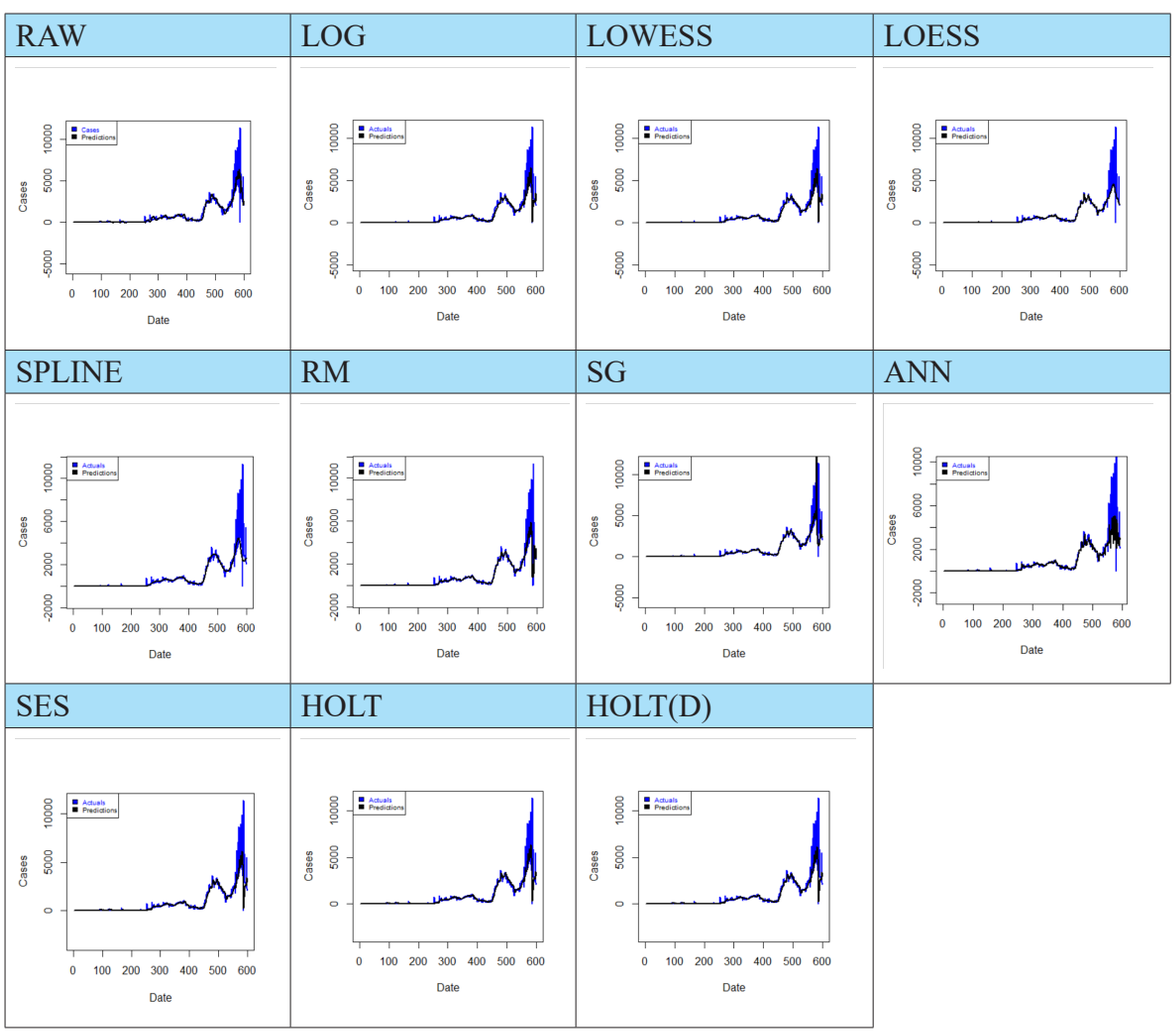


Figure 2.

The residuals of the fit of the models of COVID-19 cases.

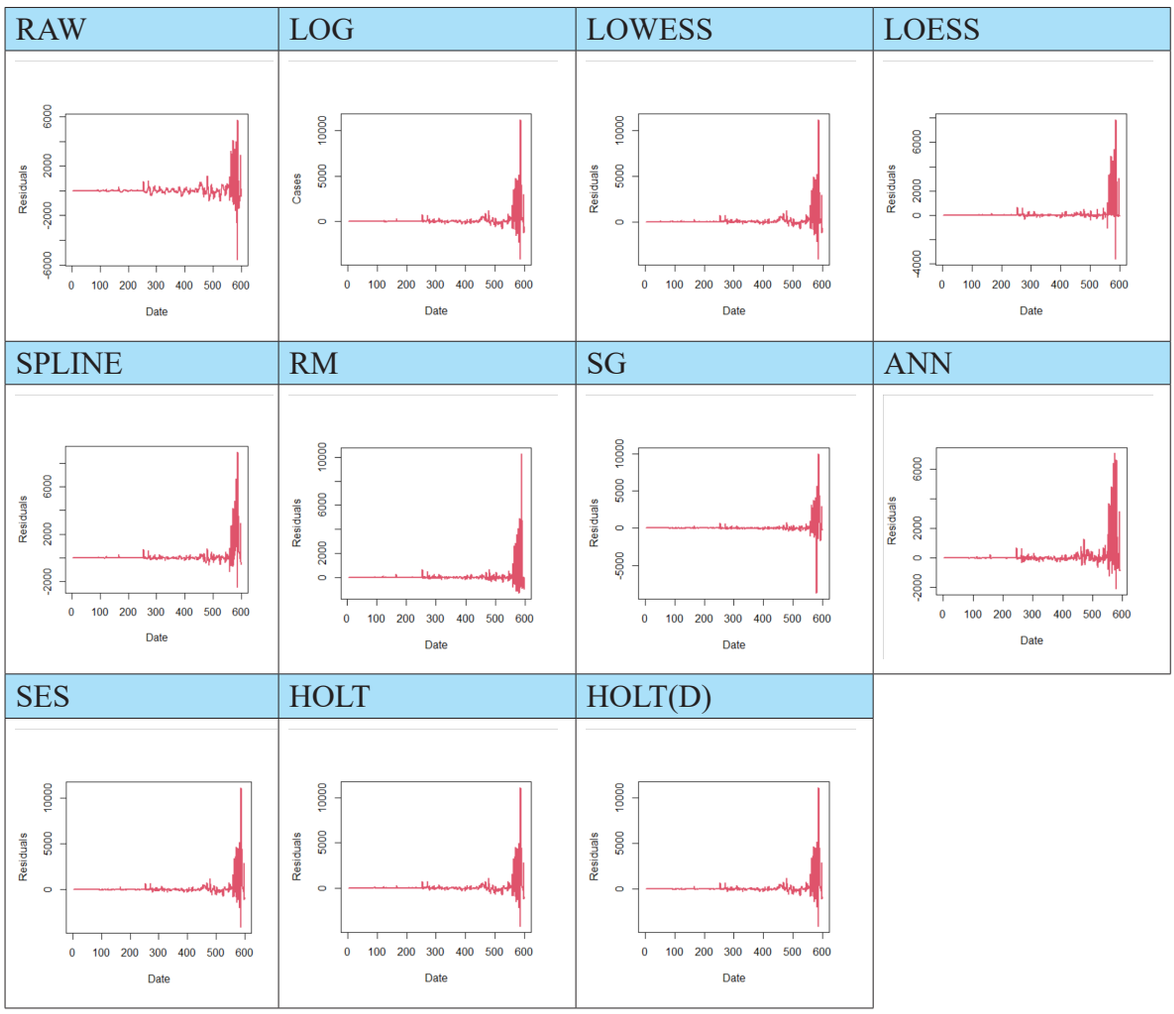


Figure 3.

Next day predictions of Covid-19 cases.

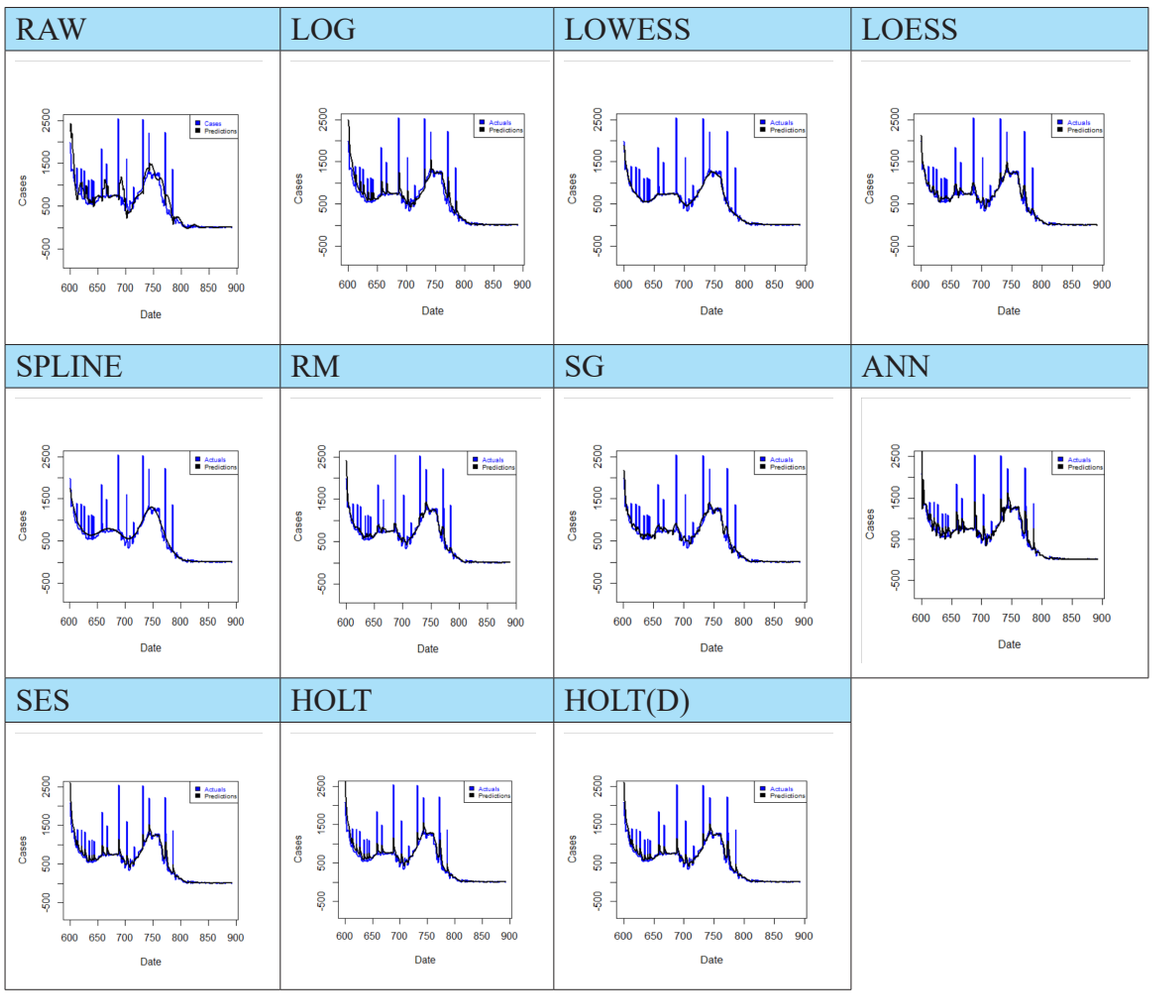


Figure 4.

Residuals of Next day predictions of Covid-19 cases.

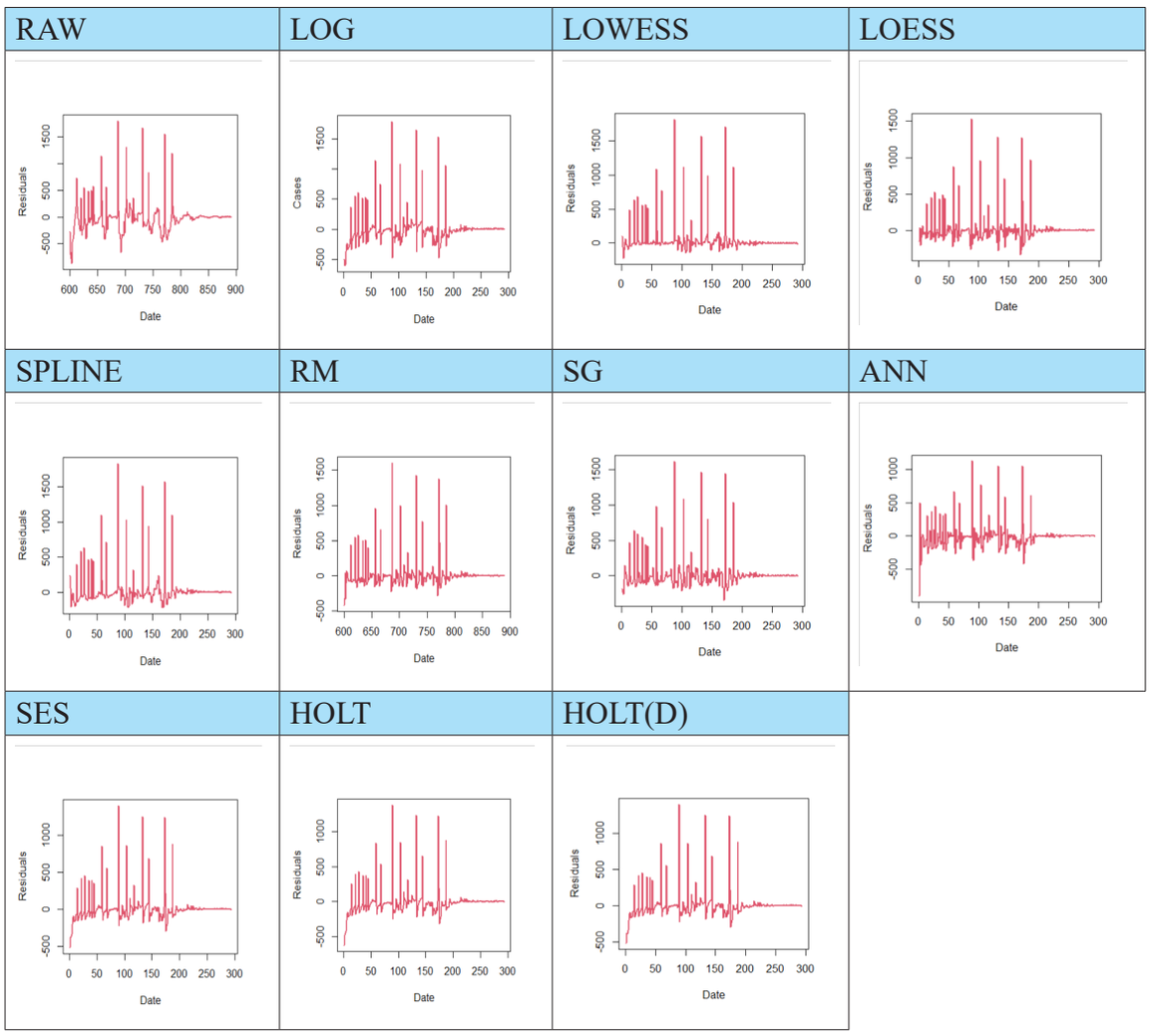


Figure 5.

One week ahead predictions of Covid-19 cases.

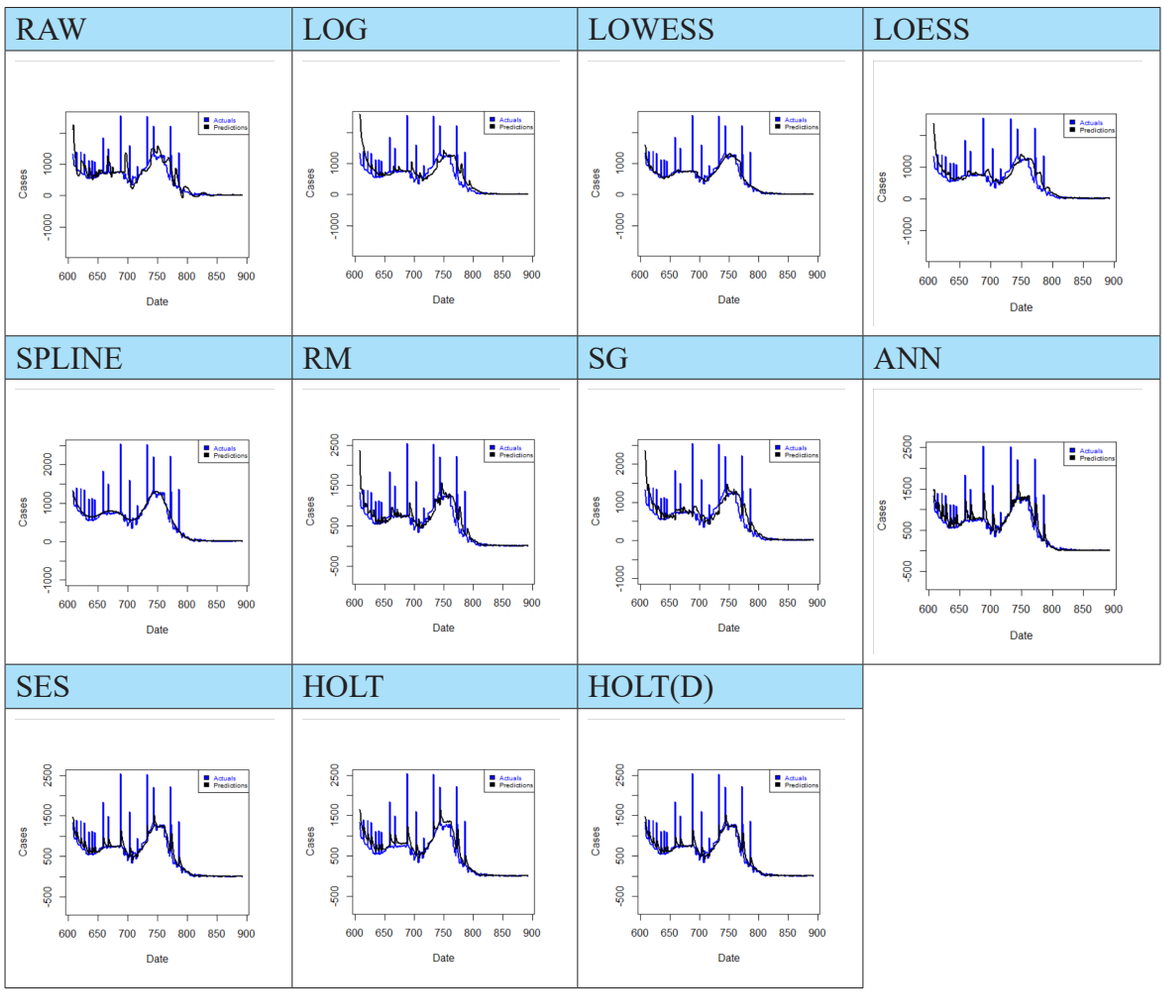


Figure 6.

The residuals of one week ahead predictions of Covid-19 cases.

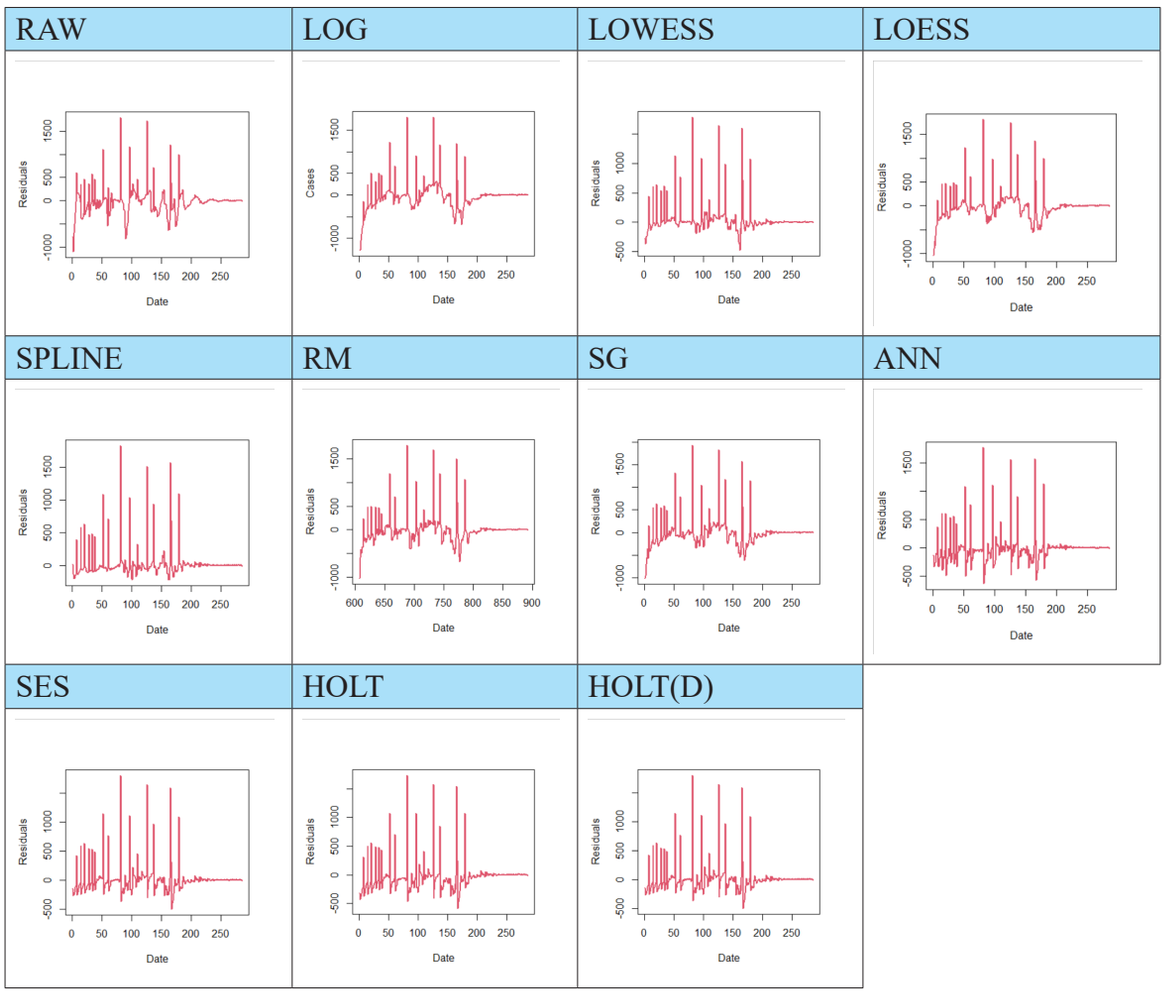


Figure 7.

One month ahead predictions of Covid-19 cases.

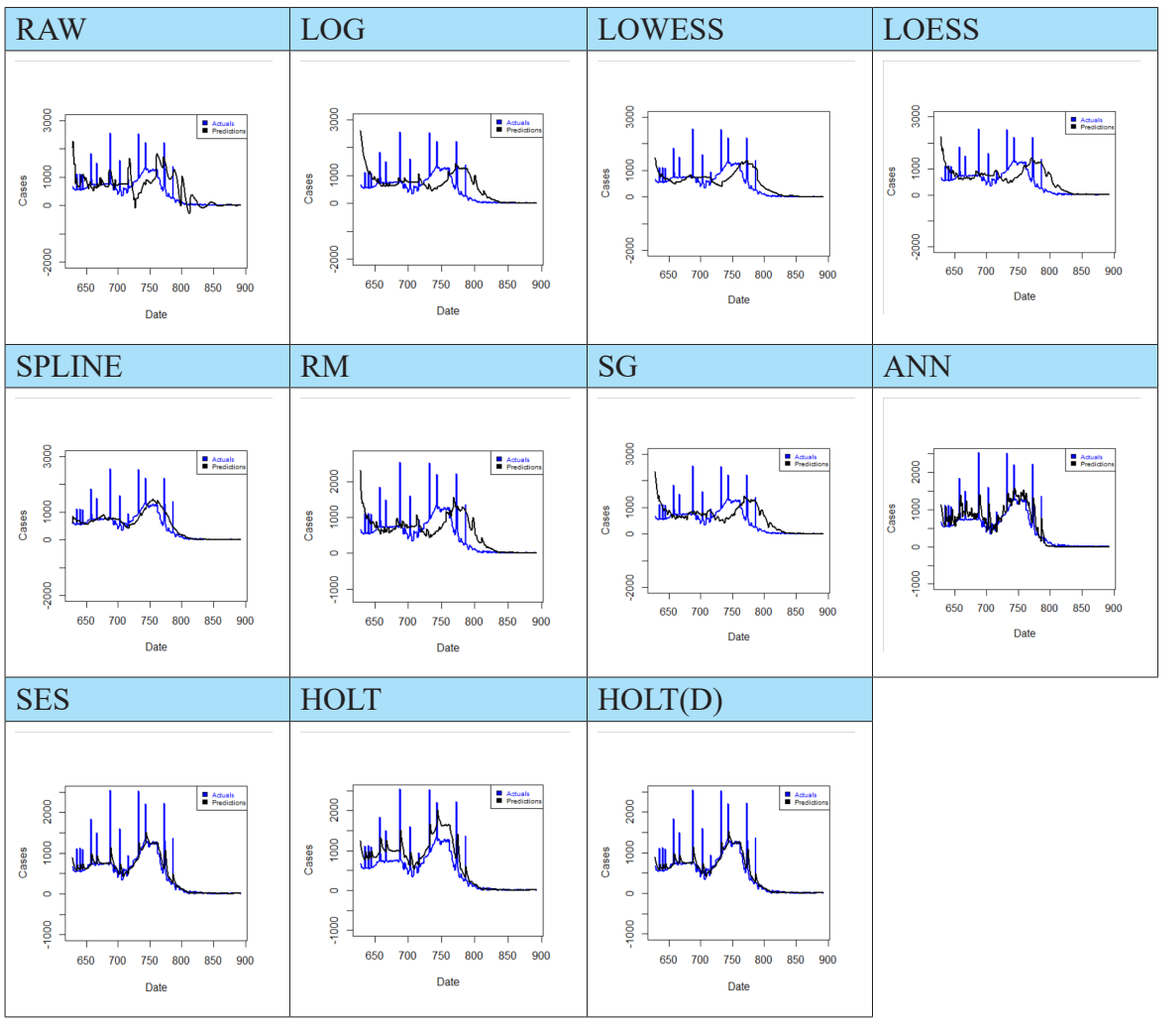


Figure 8.

Residuals of one month ahead predictions of Covid-19 cases.