Forecasting Performance Comparison of ARIMA and VAR Models in Dengue Cases Prediction for Five Districts in Sri Lanka

Dilusha N. Senarathna [1], Malima S. Atapattu[2], and Sachith P. Abeysundara[3]

[1, 2, 3] Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Peradeniya, Sri Lanka

dilusha.ns@sci.pdn.ac.lk
malima.atapattu@ sci.pdn.ac.lk
sachitha@ sci.pdn.ac.lk

ABSTRACT

Currently, there is an increase in the number of dengue cases reported in Sri Lanka. Therefore, the main objective of this study is to build a suitable time series model to predict the number of dengue cases in Sri Lanka. We considered the five districts, Colombo, Kurunegala, Nuwara Eliya, Trincomalee, and Puttalam, for the study. Also, we are interested in identifying a pattern of dengue cases in each district and studying the effect of climatic factors on dengue disease as the dengue vector mosquito lays its eggs in water. The data used in this study was the number of dengue cases reported in each district and the precipitation and temperature data as the climatic factors from 2010 January to 2022 June. Dengue data sets were collected from monthly reported dengue cases in Epidemiology Unit, Ministry of Health in Sri Lanka. Daily precipitation and temperature data were obtained from National Centers for Environmental Information which are publicly available, and monthly data was calculated. A comparative study of Vector Auto Regression (VAR) and Auto Regression Integrated Moving Average (ARIMA) in predicting dengue cases was carried out. The VAR and ARIMA models were fitted to the data predicted the dengue cases, and the accuracy by was checked by calculating Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percent Error (MAPE). Finally, the results of our study show that the VAR (1) model, which used only the climatic factor with a MAPE of 41%, is the best model to predict the dengue cases of Kurunegala. VAR (5) model with only the climatic factor precipitation with a MAPE of 51% is the best model to predict the dengue cases of Colombo, ARIMA (1,0,1) (1,0,0) [12] model with a MAPE of 40% is the best model for Nuwara Eliya. ARIMA (1,0,1) is the best model for the Puttalam and Trincomalee districts, with a 31% and 24% MAPE respectively.

1. INTRODUCTION

Time series forecasting is a technique for predicting future events by analyzing past trends based on the assumption that future trends will be similar to historical trends. Forecasting involves using models fit on historical data to predict future values. VAR models with different lag orders are used for the multivariate time series data and ARIMA models for univariate time series data. According to the National Dengue Control Unit (NDCU), dengue has a seasonal transmission in Sri Lanka, with two peaks occurring with the monsoon rains in June-July and October-December, respectively. Most cases occur during June- July, the summer one monsoon. Dengue is a flavivirus transmitted by

mosquito vectors, such as *Aedes aegypti* and *Aedes albopictus*. Mosquitoes require standing water to reproduce, and the average temperature of Sri Lanka which varies from 17◦C to 35◦C is ideal for transmission. Some of the previous research on dengue disease suggests that there might be some relationship between the number of reported dengue cases and climatic factors such as rainfall and temperature. But some of them indicate that there is no such relationship between those factors and dengue cases. However, these conclusions highly depend on the geographical area in which they conducted their research. The opposite findings and the increasing trend of dengue cases in Sri Lanka have motivated us to achieve this research. The principal objective of this research is to find the best time series model that forecasts the number of dengue cases in each of the five districts in Sri Lanka. The other objective is to investigate the relationship between dengue cases and the climatic factors in each district.

2. MATERIALS AND METHODS

The data for this study was the number of dengue cases reported in each district and the precipitation and temperature data as the climatic factors. Dengue data sets were collected from monthly reported dengue cases in the Epidemiology Unit, Ministry of Health in Sri Lanka. Daily precipitation and temperature data were obtained from National Centers for Environmental Information which are publicly available, and monthly data was calculated. The data from 2010 to 2021 were used for model construction, and the rest from January to June 2022 were used for model validation. The univariate time series model ARIMA and multivariate time series model VAR were used to predict the number of Dengue cases in each district. At first, data were checked for stationary data series dengue cases, total precipitation, and average temperature with the help of a time series graph and the Augmented Dickey-Fuller test. Then both models need to identify the optimal parameters for ARIMA (p,d,q) and VAR(p). The analysis was done using R statistical software. The *auto.Arima()* and *lagselect()* functions were used to find the optimal parameters for the ARIMA and VAR models.

3. RESULTS AND DISCUSSION

3.1. Preliminary Analysis

The dataset does not follow normal distribution according to the normal Q-Q plot and Shapiro-Wilks test. Then Kendall's rank correlation and Spearman's rank correlation were used for testing the correlation between dengue cases and climate factors.

Table 24: Kendall's rank coefficient dengue vs. total precipitation

District	Dengue vs. Total PRCP		Accept / Reject
	Correlation Coefficient	P-value	
Colombo	-0.1206	0.0309	Correlated
Kurunegala -ala	-0.1406	0.0120	Correlated
Nuwara - Eliya	-0.0584	0.3002	Uncorrelated
Trincomale e	-0.1116	0.0471	Correlated
Puttalam	-0.0688	0.2192	Uncorrelated

Table 25: Kendall's rank coefficient dengue vs. average temperature

District	Dengue vs. average temperature		Accept / Reject
	Correlation Coefficient	P-value	
Colombo	-0.0478	0.3928	Uncorrelated
Kurunegala-a	-0.0149	0.7905	Uncorrelated
Nuwara -Eliya	-0.1195	0.0343	Correlated
Trincomalee-e	-0.0711	0.2055	Uncorrelated
Puttalam	-0.1696	0.0025	Correlated

Table 26: Spearman's rank coefficient dengue vs. total precipitation

District	Dengue vs. Total PRCP		Accept / Reject
	Correlation Coefficient	P-value	
Colombo	-0.1832	0.0269	Correlated
Kurunegala	-0.2042	0.0134	Correlated
Nuwara Eliya	-0.0937	0.2608	Uncorrelated
Trincomalee	-0.1641	0.0478	Correlated
Puttalam	-0.1011	0.2249	Uncorrelated

Table 27: Spearman's rank coefficient dengue vs. average temperature

District	Dengue vs. Total PRCP		Accept / Reject
	Correlation Coefficient	P-value	
Colombo	-0.0639	0.4437	Uncorrelated
Kurunegala	-0.2014	0.8094	Uncorrelated
Nuwara Eliya	-0.1918	0.0204	Correlated
Trincomalee	-0.1113	0.1809	Uncorrelated
Puttalam	-0.2460	0.0028	Correlated

3.2. Model fitting

Researchers have found a direct relationship between precipitation and the increase of dengue cases. According to Edirisinghe (2017) findings, the numbers of dengue cases reported in June, July, January, and February are the highest annually, irrespective of the geographical locations of the country. Therefore, six models were fitted with ARIMA*(p,d,q)*, three VAR models with both total precipitation and average temperature with lag orders *1*,*2*, and *p*, and another three VAR models only considering the effect of total precipitation with lag orders *1*,*2*, and *p*.

Table 28: Fitted ARIMA models

District	ARIMA
Colombo	ARIMA (1,0,0) (1,0,0) [12]
Kurunegala	ARIMA (5,0,1) (1,0,0) [12]
Nuwara Eliya	ARIMA (1,0,1) (1,0,0) [12]
Puttalam	ARIMA (1,0,1)
Trincomalee	ARIMA (1,0,1)

Table 29: Fitted VAR models

District	VAR with both climate factors (lag order)	VAR with only total precipitation (lag order)
Colombo	1,2,5	1,2,9
Kurunegala	1,2,5	1,2,5
Nuwara Eliya	1,2,7	1,2,3
Trincomalee	1,2,10	1,2,3
Puttalam	1,2,5	1,2

For each district, at least six models were tested, and the conclusions are made based on the tested models. Tables 5 and 6 contain all the tested models.

3.3. Model selection

The number of dengue cases for January to June 2022 was predicted using fitted models. The Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) were used to check the accuracy of the fitted models and then to select the best fit model to predict the number of dengue cases for each district.

In table A1, A2, A3, A4, and A5, the error measurements were included for the tested models for each district. The best model was selected depending on the minimum error measurements for at least two measurements of

tested models and with several modifications and advancements, these models can be changed.

The following are the graphs of forecasted data using the selected model and actual data and future predictions for the next two years for each district.

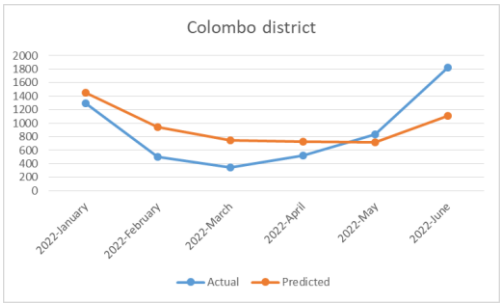

Figure 44: Forecasted values for Colombo district

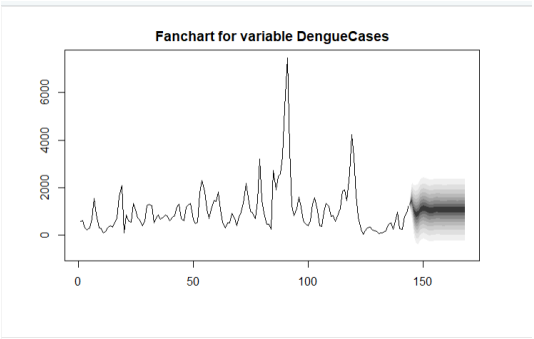

Figure 45: Forecasted values for the next two years of Colombo district

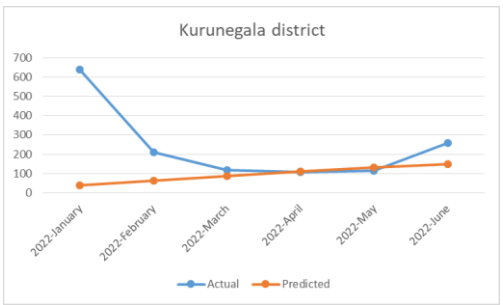

Figure 46: Forecasted values for Kurunegala district

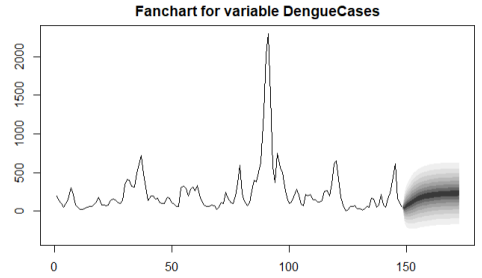

Figure 47: Forecasted values for the next two years of Kurunegala district

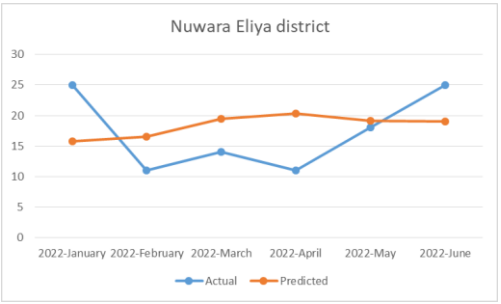

Figure 48: Forecasted values for Nuwara Eliya district

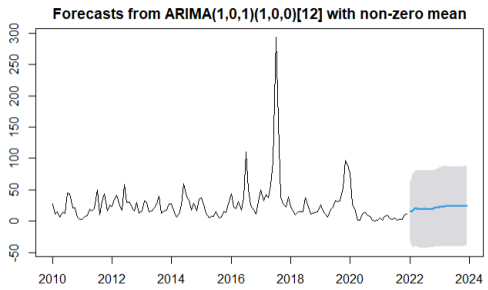

Figure 49: Forecasted values for the next two years of Nuwara Eliya district

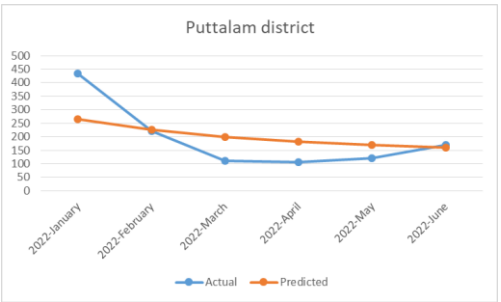

Figure 50: Forecasted values for Puttalam district

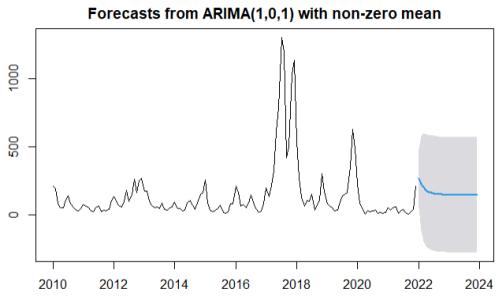

Figure 51: Forecasted values for the next two years of Puttalam district

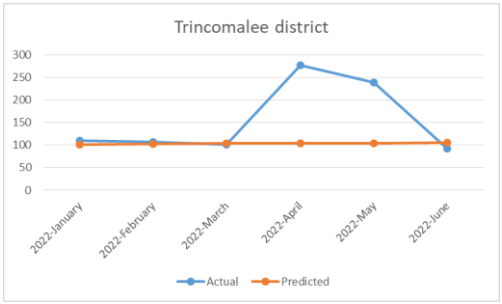

Figure 52: Forecasted values for Trincomalee district

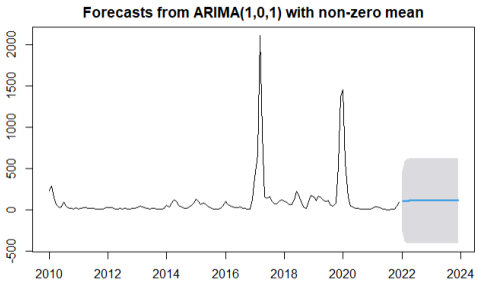

Figure 53: Forecasted values for the next two years of Trincomalee district

4. CONCLUSION

The forecasting comparison of ARIMA and VAR models shows that if the variables are strongly correlated for a particular district, the VAR model will give the most accurate predictions. When modeling with climatic factors that are not highly correlated, there is no noticeable difference in the performances of ARIMA and VAR models. Finally, in our study, we found that for Colombo district VAR (5) model with total precipitation, for the Kurunegala district VAR (1) model with total precipitation, for Nuwara Eliya district ARIMA (1,0,1) (1,0,0) [12], for Puttalam and Trincomalee districts ARIMA (1,0,1) models can be selected as the best models according to the above accuracy measurements. Also, according to these selected final models, the monthly average temperature will not affect the number of dengue cases in five considered districts in Sri Lanka. Results suggested that paying attention to conducting dengue prevention and controlling programs in the Colombo district is important. This study can be further improved by considering other environmental factors that affect the number of dengue cases in Sri Lanka with VAR models. Also, this can be extended to the other districts in Sri Lanka.

REFERENCES

Edirisinghe, G. (2017). Contribution of rainfall patterns for an increased dengue epidemic in Sri Lanka. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS) ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

N.D.B. Ehelepola, Kusalika Ariyaratne, W. M. N. B. S. R. N. W. (2015). A study of the correlation between dengue and weather in Kandy city, Sri Lanka (2003 -2012) and lessons learned. Infectious Diseases of Poverty (2015) 4:42 DOI 10.1186/s40249-015-0075-8.

Parvin, S. and Khanam, M. (2018). Comparison between Arima and var model regarding the forecasting of the price of jute goods in Bangladesh. Dhaka Univ. J. Sci. 66(2): 91-94, 2018 (July). Schwabacher, M., & Goebel, K. F. (2007).

Khan, M. S. and Khan, U. (2020). Comparison of forecasting performance with var vs. arima models using economic variables of Bangladesh. Asian Journal of Probability and Statistics 10(2): 33-47, 2020; Article no. AJPAS.63149 ISSN: 2582-0230.

Deshmukh, S. S. and Paramasivam, R. (2016). Forecasting of milk production in India with arima and var time series models. Asian J. Dairy Food Res, 35(1) 2016: 17-22.

Gayan P. Withanage, Sameera D. Viswakula, Y. S. G. M. H. (2018). A forecasting model for dengue incidence in the district of Gampaha, Sri Lanka.

Goswami, S. (2014). Study of effectiveness of time series modeling (arima) in forecasting stock prices. International Journal of Computer Science Engineering and Applications.

APPENDIX

VAR1 – VAR model with both climate factor and temperature

VAR2 – VAR model with only climate factor

Table A1: Error measurements of Colombo District

Error Measurement		Lag =1		Lag =2		Lag=5	
	ARIMA	VAR1	VAR2	VAR1	VAR2	VAR1	VAR2
RMSE	506.6072	537.0368	554.524	557.6992	588.3101	399.3010	**389.3298**
MAE	445.0541	490.5952	507.8715	506.452	534.9025	350.3069	**340.8404**
MAPE	70.2830	81.6529	86.3769	85.6656	92.0595	56.7515	**51.3436**

Table A2: Error measurements of Kurunegala District

Error Measurement		Lag =1		Lag =2		Lag =5	
	ARIMA	VAR1	VAR2	VAR1	VAR2	VAR1	VAR2
RMSE	229.1245	256.9124	253.6411	271.6045	258.6238	**223.9638**	235.8672
MAE	212.2675	151.6317	**149.4061**	167.1987	156.033	170.9449	187.8177
MAPE	149.3748	41.9245	**41.6299**	58.3314	53.8747	90.0517	106.9178

Table A3: Error measurements of Nuwara Eliya District

Error Measurement		Lag =1		Lag =2		Lag =7	Lag =3
	ARIMA	VAR1	VAR2	VAR1	VAR2	VAR1	VAR2
RMSE	**6.7178**	9.0828	8.4597	15.1626	11.5796	25.5136	12.7408
MAE	**6.1276**	8.1396	7.5468	13.5701	10.5577	22.5049	11.2072
MAPE	**40.2804**	59.3167	53.4356	94.9026	71.9998	162.5242	82.7065

Table A4: Error measurements of Puttalam District

Error Measurement		Lag = 1		Lag = 2		Lag =10	Lag = 3
	ARIMA	VAR1	VAR2	VAR1	VAR2	VAR1	VAR2
RMSE	**85.8027**	148.1345	435.8258	160.9142	162.4876	154.6141	146.6201
MAE	**65.4920**	90.0798	258.2348	108.1175	94.8088	121.7522	90.1872
MAPE	39.1985	33.0665	168.5866	47.3234	**31.9705**	63.9549	34.9652

Table A5: Error measurements of Trincomalee District

Error Measurement		Lag = 1		Lag =2	
	ARIMA	VAR1	VAR2	VAR1	VAR2
RMSE	89.9479	94.4568	**88.6405**	114.7482	93.8781
MAE	**55.9922**	60.0952	59.8489	77.7555	63.5639
MAPE	**24.3755**	26.5185	28.5869	37.3337	30.3140