



Prediction of Dengue Cases

A Time Series Approach

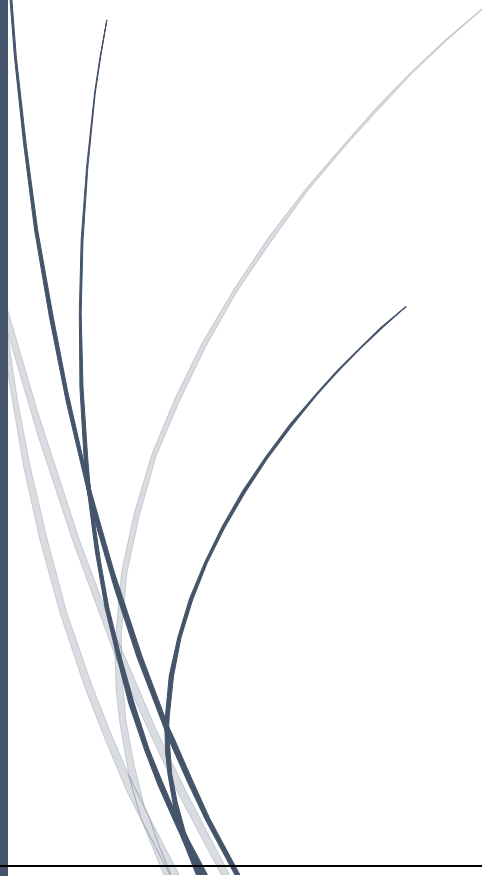
Submitted By: -

Aravind S (A0163301X)

Kavya AK (A0163250R)

Praman Shukla (A0163239A)

Saravanan kalastha sekar (A0163309H)



Objective:

To predict the relationship between the total Dengue cases recorded in various regions of Singapore and the amount of rainfall as measured in Singapore Changi climate station for the years 2012-2017.

Time Series Data Description:**Total Monthly Rainfall.**

Singapore is situated near the equator and has a typically tropical climate, with abundant rainfall, high and uniform temperatures, and high humidity all year round. Singapore's climate is characterized by two monsoon seasons separated by inter-monsoonal periods. The Northeast Monsoon occurs from December to early March, and the Southwest Monsoon from June to September. Rainfall is plentiful in Singapore and it rains an average of 178 days of the year as much of the rain is heavy and accompanied by thunder. The long-term mean annual rainfall total is 2328.7mm. While there is no distinct wet or dry season in Singapore, monthly variations in rainfall do exist. Higher rainfall occurs from November to January during the wet phase of Northeast Monsoon season whereas the driest month is February which is during the dry phase of the Northeast Monsoon when the rain-belt has moved further south to affect Java.

This time series is considered as the independent variable(X_t) and the essential data set is extracted from the data.gov.sg site. The time series data consists of total monthly rainfall recorded (measured in mm) in Singapore for the years 2012-2017.

Data Source: <https://data.gov.sg/dataset/rainfall-monthly-total>

Monthly Dengue Cases:

Dengue fever and dengue hemorrhagic fever (a more severe form) are the most common mosquito-borne viral diseases in Singapore. Dengue fever is an illness caused by infection with a virus transmitted by the Aedes mosquito, which is not contagious and does not spread directly from person to person. A mosquito is infected when it takes a blood meal from a dengue-infected person and later transmits the virus to other people they bite.

This time series is considered as the dependent variable(Y_t) and the required data set is extracted from the data.gov.sg site. The times series data set contains of various Infectious Disease cases recorded, where only the dengue disease cases for the years 2012-2017 are considered for the modelling.

Data Source: <https://data.gov.sg/dataset/weekly-infectious-disease-bulletin-cases>

Null Hypothesis:

The null hypothesis states that “No relationship is present between the two given times series data of rainfall and dengue”.

- Null Hypothesis is retained when no Correlation is experienced.
- Null Hypothesis is rejected when Correlation is observed.

Modelling:

The monthly rainfall data is taken as the independent variable ‘X’ and the monthly recorded cases of dengue is taken as the dependent variable ‘Y’ and time series modelling is performed in JMP.

Checking Stationarity:

The input series is checked for the stationarity. When looking at the single mean ADF and Trend ADF, the values are above the critical value mentioned in the Augmented-Dickey Fuller testing criterion which implies that data itself attained the stationarity at the beginning. Further on, we decided to build the model with the input data series which is the monthly rainfall.

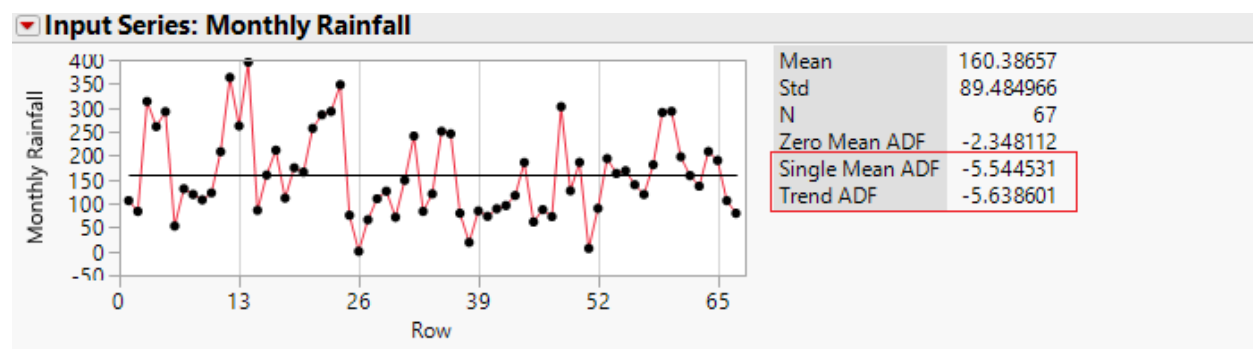


Figure 1

In general, rainfall data has a seasonality associated with it. The data that we took is the Singapore rainfall data, though Singapore climate didn't have a regular pattern in terms of rainfall over a year, but from the time basis characteristics graph, it has been noted that there are some similarities in the pattern of data distribution in the timeframe of 12 periods which encourages us to take the seasonality observation as 12 to build our model.

Correspondingly, we have built a SARIMA (Seasonal Auto Regressive Integrated Moving Average) Model for this data because of this seasonality component associated with the input series.

Seasonal ARIMA:

The seasonal part of an ARIMA model has the same structure as the non-seasonal part but it may have an AR factor, an MA factor, and/or an order of differencing. In the seasonal part of the model, these factors operate across multiples of lags (the number of periods in a season).

A seasonal ARIMA model is classified as an **ARIMA(p,d,q)x(P,D,Q)** model

Where,

p-number of autoregressive terms

d-number of non-seasonal differences needed for stationarity

q-number of lagged forecast errors in the prediction equation.

P-number of seasonal autoregressive (SAR) terms

D-number of seasonal differences

Q-number of seasonal moving average (SMA) terms.

We tried modelling the input data series with various combinations of SARIMA to find the best model which achieves the white noise and all the parameters being significant. For accomplishing this, various models are performed with different **(p,d,q)(P,D,Q)** and the best fit model has been selected based on AICC and BIC criteria.

From the model comparison table, we observed the model with (1,1,1)(0,1,1) has the lowest AICC and BIC value achieves the white noise with significant parameters which has been highlighted in the below figures 2& 3.

Model Comparison																
	Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2	.4	.6	.8	MAPE	MAE
▼	✓	☐	Seasonal ARIMA(1, 1, 1)(0, 1, 1)12	50	5646.1123	648.65932	656.61525	-0.39	640.65932	0.770937					2180.2307	70.123918
▼	✓	☐	Seasonal ARIMA(0, 1, 1)(0, 1, 1)12	51	6478.1612	652.24946	658.21641	-0.55	646.24946	0.128065					2650.7609	77.493250
▼	✓	☐	Seasonal ARIMA(1, 0, 0)(1, 1, 1)12	51	4921.8831	653.85434	661.88367	-0.17	645.85434	0.057403					1469.7663	69.539577
▼	✓	☐	Seasonal ARIMA(1, 1, 0)(1, 1, 1)12	50	6204.1482	656.00834	663.96428	-0.59	648.00834	0.019552					1407.0285	78.043594
▼	✓	☐	Seasonal ARIMA(1, 1, 0)(0, 1, 1)12	51	7372.2124	657.56641	663.53336	-0.57	651.56641	0.008972					1739.2866	78.009487
▼	✓	☐	Seasonal ARIMA(1, 1, 1)(1, 1, 0)12	50	8875.5752	657.82653	665.78246	-0.57	649.82653	0.007877					1849.2007	80.104203
▼	✓	☐	Seasonal ARIMA(1, 1, 1)(1, 1, 1)12	49	6310.491	658.03153	667.97645	-0.60	648.03153	0.007110					1368.5997	78.655198
▼	✓	☐	Seasonal ARIMA(1, 0, 1)(1, 1, 0)12	51	9176.1132	667.06653	675.09586	-0.36	659.06653	0.000078					1596.9915	79.653760
▼	✓	☐	Seasonal ARIMA(1, 1, 1)(0, 1, 0)12	51	13253.033	672.12187	678.08882	-0.92	666.12187	0.000006					3444.8065	84.904280
▼	✓	☐	Seasonal ARIMA(1, 1, 1)(0, 0, 1)12	62	7418.0739	782.83826	791.59688	0.073	774.83826	0.000000					1390.8453	64.888662
▼	✓	☐	Seasonal ARIMA(0, 1, 1)(1, 0, 1)12	62	7272.7785	790.00566	798.76428	-0.04	782.00566	0.000000					1890.0047	73.421975
▼	✓	☐	Seasonal ARIMA(1, 0, 1)(1, 0, 1)12	62	6596.078	791.69277	802.71624	0.142	781.69277	0.000000					1299.8737	64.571326

Figure 2

Model: Seasonal ARIMA(1, 1, 1)(0, 1, 1)12									
Model Summary									
DF			50	Stable	Yes				
Sum of Squared Errors			282305.617	Invertible	Yes				
Variance Estimate			5646.11234						
Standard Deviation			75.1406171						
Akaike's 'A' Information Criterion			648.659316						
Schwarz's Bayesian Criterion			656.615252						
RSquare			-0.388409						
RSquare Adj			-0.4717135						
MAPE			2180.23075						
MAE			70.1239179						
-2LogLikelihood			640.659316						
Parameter Estimates									
Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob> t	Constant	Mu	
AR1,1	1	1	0.3364183	0.1349639	2.49	0.0160*	Estimate	1.21154394	
MA1,1	1	1	0.9999537	0.1072703	9.32	<.0001*	0.8039584		
MA2,12	2	12	0.9999695	0.2374623	4.21	0.0001*			
Intercept	1	0	1.2115439	0.9459175	1.28	0.2062			

Figure 3

When examining the parameter estimates, it has been evident that the model terms are significant and thus we can proceed with the model. The next entity which we considered is the residual distribution. It attained the homoscedastic state, and moreover we look at the ACF and PACF plots. All the lag spikes are within the confidence interval of 95% and the P-values are insignificant for the autocorrelation lags, which means the model has eliminated all the autocorrelation among the residuals and in the time series. Refer fig. 4&5

The residuals left over after fitting the model should be white noise and from the plot it has been clear that it has achieved the white noise. And ACF and PACF of the residuals show no significant autocorrelations or partial autocorrelations.

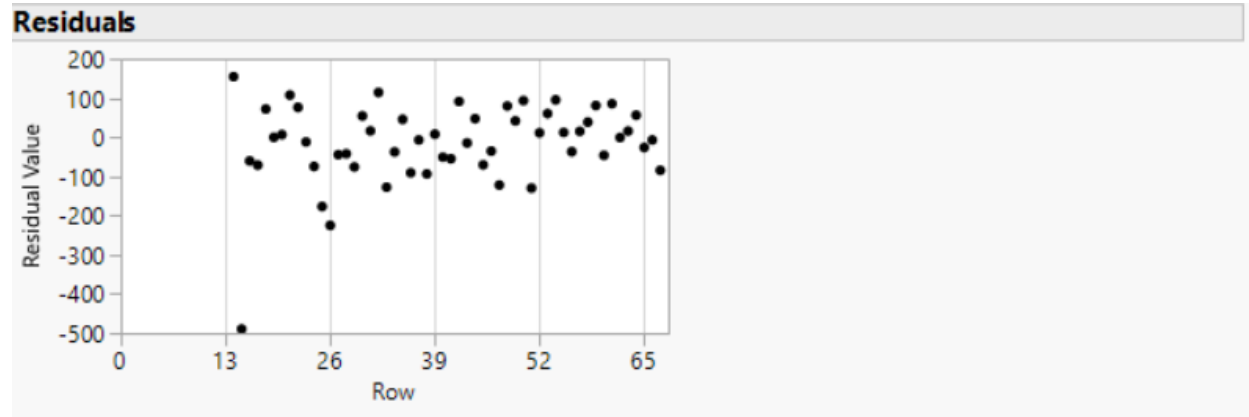


Figure 4

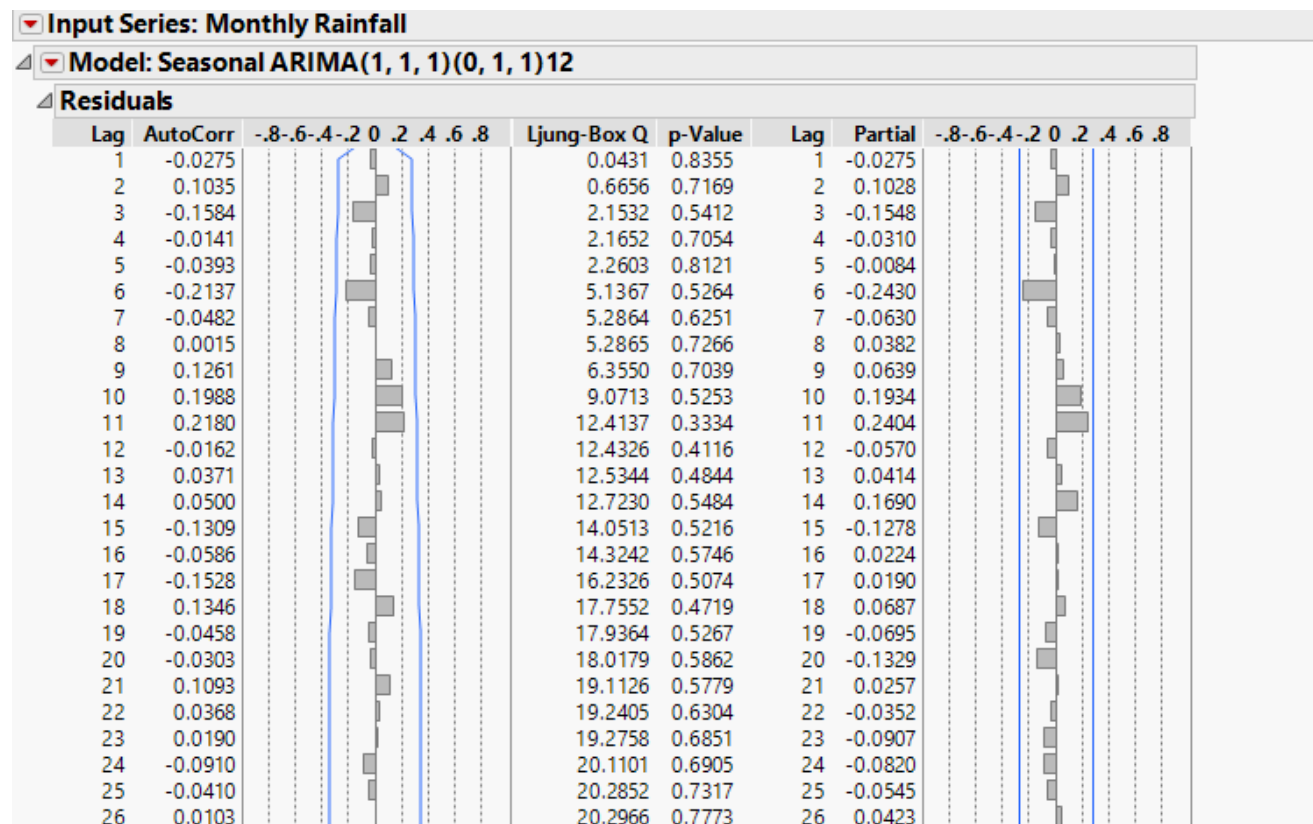


Figure 5

Seasonal ARIMA in Output Series:

The same Seasonal ARIMA model has been implemented on the output series to check whether it attains the white noise. On running the time series data with the SARIMA model, by observing at the output ACF and PACF, we infer that there are no significant auto correlations and partial autocorrelations (With reference to the P-value) Ref fig.7. This infers this model has been complementing with the output series and we can proceed with this model to see the pre-whitening plot.

Time Series Monthly Dengue Cases

Model: Seasonal ARIMA(1, 1, 1)(0, 1, 1)12

Model Summary

Sum of Squared Errors	8052304.99	Invertible	Yes
Variance Estimate	161046.1		
Standard Deviation	401.305494		
Akaike's 'A' Information Criterion	825.167875		
Schwarz's Bayesian Criterion	833.123811		
RSquare	0.46267309		
RSquare Adj	0.43043348		
MAPE	31.6950234		
MAE	364.366114		
-2LogLikelihood	817.167875		

Figure 6

Time Series Monthly Dengue Cases

Model: Seasonal ARIMA(1, 1, 1)(0, 1, 1)12

Residuals

Lag	AutoCorr		Ljung-Box Q	p-Value	Lag	Partial	
0	1.0000				0	1.0000	
1	-0.0450		0.1156	0.7338	1	-0.0450	
2	-0.0759		0.4505	0.7983	2	-0.0781	
3	-0.1615		1.9961	0.5732	3	-0.1700	
4	-0.1319		3.0477	0.5499	4	-0.1620	
5	-0.0357		3.1263	0.6805	5	-0.0902	
6	-0.0011		3.1264	0.7928	6	-0.0728	
7	0.1186		4.0307	0.7762	7	0.0514	
8	0.0805		4.4572	0.8137	8	0.0504	
9	-0.0225		4.4914	0.8762	9	-0.0210	
10	-0.1500		6.0373	0.8121	10	-0.1310	
11	-0.1376		7.3684	0.7685	11	-0.1377	
12	-0.2442		11.6625	0.4732	12	-0.3178	
13	0.2169		15.1334	0.2991	13	0.0875	
14	0.1686		17.2828	0.2414	14	0.0697	
15	0.0117		17.2933	0.3016	15	-0.0896	
16	0.0167		17.3155	0.3655	16	-0.0150	
17	-0.0578		17.5884	0.4152	17	-0.0100	
18	-0.0306		17.6672	0.4778	18	-0.0002	
19	-0.0397		17.8033	0.5356	19	0.0154	
20	0.1478		19.7450	0.4740	20	0.1530	
21	0.0365		19.8672	0.5297	21	-0.0301	
22	-0.0311		19.9585	0.5856	22	-0.1187	
23	-0.1304		21.6170	0.5435	23	-0.1636	
24	-0.1435		23.6921	0.4793	24	-0.2132	
25	-0.0551		24.0083	0.5189	25	-0.0345	

Figure 7

Pre-whitening:

Pre-whitening is performed to find the cross-correlation between the dengue cases registered and the monthly rainfall amount.

From the figure 8, we can see the relationship between the variables from the input lag 1 and ranging up to 6 lags with a peak at the 3rd lag and been in the positive side of the lags. This pre-whitening plot infers there is a cross-correlation with the input and output. To determine the exact relationship between the two series, Transfer function model is performed.

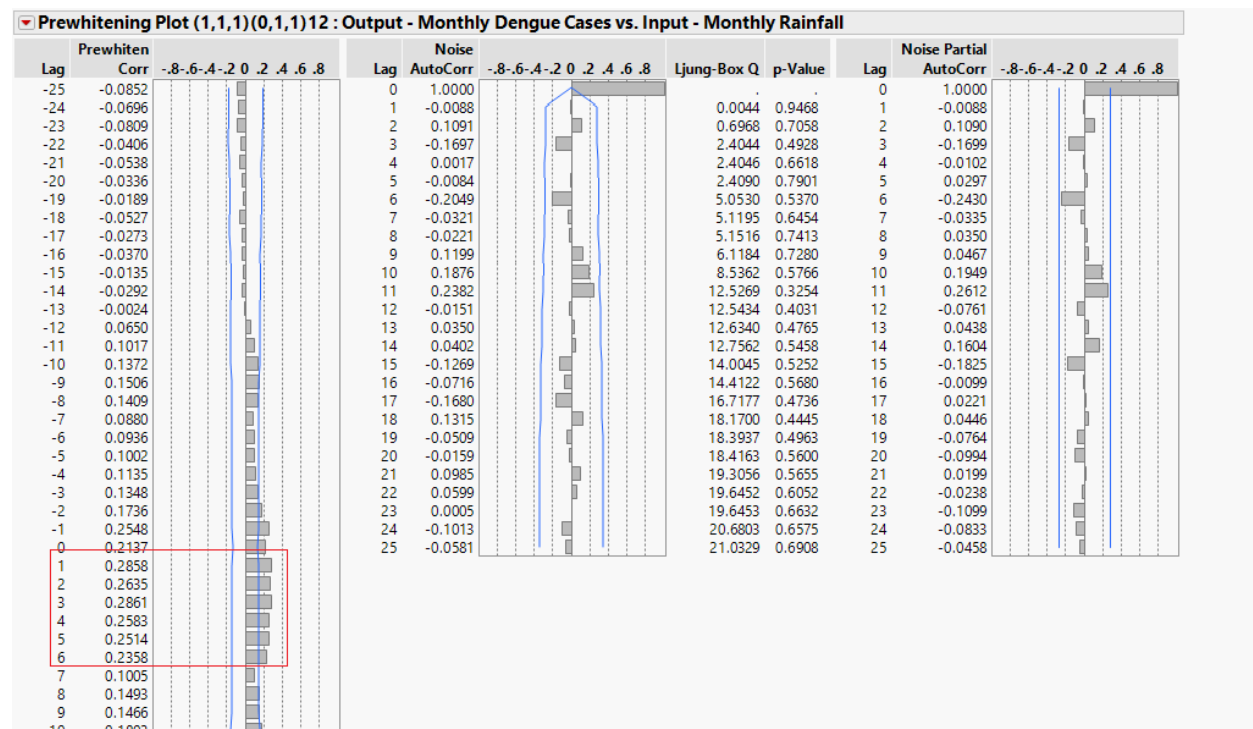


Figure 8

Transfer Function Analysis:

Transfer function method is a dynamic regression model which allows the explanatory variable to be included. The main objective of the model is to predict what happens to the forecast variable or output time series, called Y_t , if the explanatory variable or input time series, called x_t changes. Let X_t and Y_t represent input and output data of transfer function model respectively. The transfer function for the dengue cases and the amount of rainfall represented in a monthly time series has been deduced.

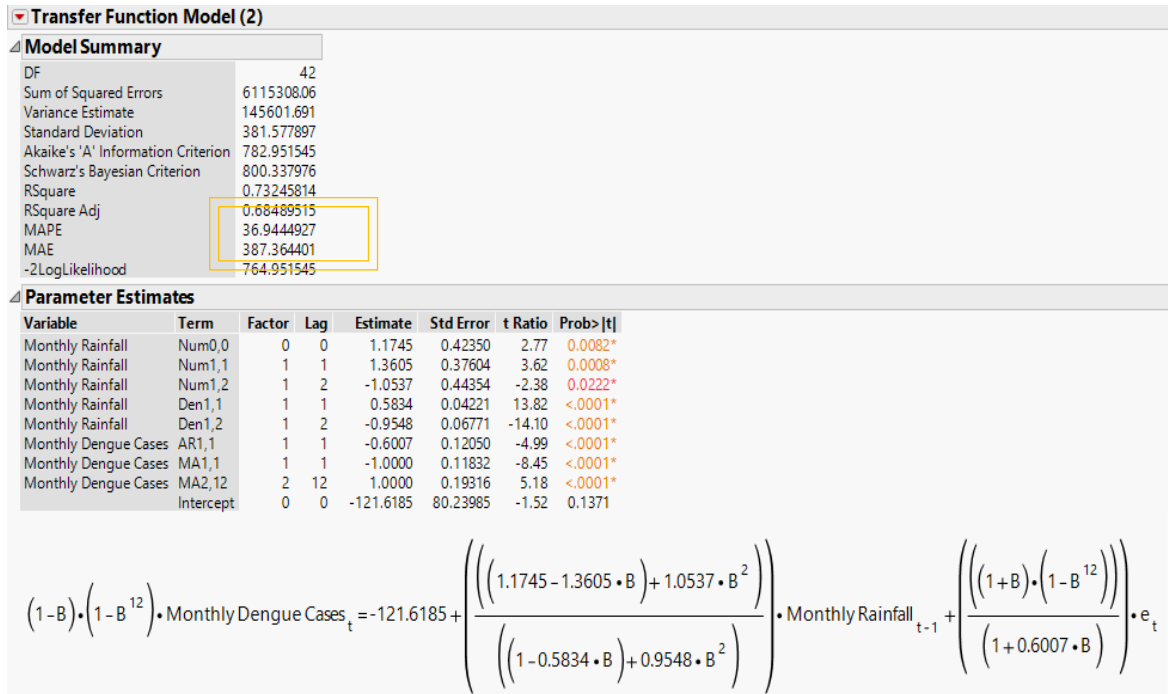


Figure 9

From the figure 9, we can see that MAPE is 36.94 and all the parameter estimates are significant. And the residuals (Ref. Fig.9) are also normally distributed and most of the lags are insignificant and all lies within the confidence interval of 95% which gives a good sense of best transfer function model to the prediction problem of dengue cases.

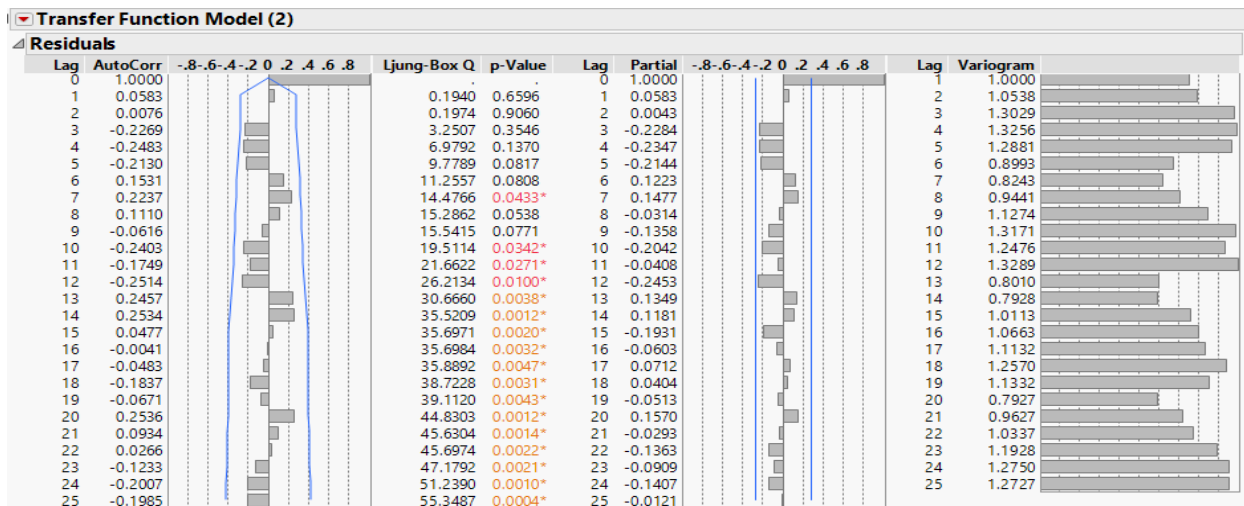


Figure 10

Equation Expansion:

The relationship between the dengue cases registered and the rainfall has been explained by the equation below which is the extended form of Transfer series equation that we obtain. The response variable, Dengue cases registered, Y_t , is having relationship with 4 lags of Explanatory Variable. X_t , Monthly rainfall and Y_t has a momentum and it's correlated within the time series.

$$Y_t = -121.6185 + [0.9827Y_{t-1} + 0.5871Y_{t-2} - 0.0309Y_{t-3} - 0.5735Y_{t-4} - Y_{t-12} + 0.9827Y_{t-14} + 0.0309Y_{t-15} + 0.5735Y_{t-16}] +$$

$$[1.1745X_{t-1} - 0.6550X_{t-2} + 0.2365X_{t-3} + 0.6329X_{t-4}] +$$

$$[1 + 0.4166e_{t-1} + 0.3714e_{t-2} + 0.9548e_{t-3} - 0.4166e_{t-13} - 0.3714e_{t-14} - 0.9548e_{t-15}]$$

The coefficients of the relational variables are:

$$X_{t-1} = 1.1745 \quad Y_{t-1} = 0.9827 \quad Y_{t-12} = -1$$

$$X_{t-2} = -0.6550 \quad Y_{t-2} = 0.5871 \quad Y_{t-14} = 0.9827$$

$$X_{t-3} = 0.2365 \quad Y_{t-3} = -0.0309 \quad Y_{t-15} = 0.0309$$

$$X_{t-4} = 0.6329 \quad Y_{t-4} = -0.5735 \quad Y_{t-16} = 0.5735$$

Inference:

This paper deals with the prediction of Dengue cases with the time series data of rainfall and cases registered. From the transfer function expansion, it's purely evident that rainfall is a major cause of dengue. As the Aedes mosquito breeds in the water-logged areas, the rainfall acts as a catalyst for it. The equation represents that the cases registered is due to the rainfall that occurred in the last 4 months where the instant last month has a very high coefficient (1.1745) which means the amount of rainfall in the previous month could be a predictor of number of dengue cases in the forthcoming month.

Thus, the null hypothesis has been rejected. Moreover, the number of Dengue cases registered has relationship with them, there will be a direct impact on the cases with consideration of the previous 4-month cases. So, it has been inferred that dengue cases registered is having a seasonality associated with it and

there's an acute increase in the number of cases at selective months in a year which has also been represented in the equation. The two inferences are that amount of rainfall is a causal variable of dengue fever and dengue cases has a cycle associated with it.

Conclusion:

A significant relationship has been identified between the monthly amount of rainfall and dengue cases registered and null hypothesis has been rejected. Therefore, for prediction of dengue cases it's essential to know about the rainfall which directly leads to water logging and cause the mosquito breeding. It's important to maintain places from water logging to avoid the dengue fever.