

# Assignment Coversheet

Faculty of Science and Engineering



MACQUARIE  
University

UNIT NO:		UNIT NAME:	
STAT826			
FAMILY NAME:		FIRST NAME:	
Palermo		Francesco	
STUDENT NO:	CONTACT PHONE NO:	TUTOR'S NAME:	
45539669			
ASSIGNMENT TITLE:		TUTORIAL/PRAC DAY:	TUTORIAL/PRAC TIME:
STAT 826 Assignment		Wednesday	Wednesday
DATE RECEIVED:	SUBMITTED WORD COUNT:	TURNITIN NUMBER:	DUE DATE:
			25/10/2019
<input type="checkbox"/> EXTENSION GRANTED?		<input type="checkbox"/> DOCUMENT ATTACHED?	

## STUDENT DECLARATION:

I certify that:

- This assignment is my own work, based on my personal study and/or research
- I have acknowledged all material and sources used in the preparation of this assignment, including any material generated in the course of my employment
- If this assignment was based on collaborative preparatory work, as approved by the teachers of the unit, I have not submitted substantially the same final version of any material as another student
- Neither the assignment, nor substantial parts of it, have been previously submitted for assessment in this or any other institution
- I have not copied in part, or in whole, or otherwise plagiarised the work of other students
- I have read and I understand the criteria used for assessment.
- The use of any material in this assignment does not infringe the intellectual property / copyright of a third party
- I have kept a copy of my assignment and this coversheet
- I understand that this assignment may undergo electronic detection for plagiarism, and a copy of the assignment may be retained on the database and used to make comparisons with other assignments in future. To ensure confidentiality all personal details would be removed
- I have read and understood the information on plagiarism. For the University's policy in full, please refer to [mq.edu.au/academic/honesty](http://mq.edu.au/academic/honesty) or Student Information in the Handbook of Undergraduate Studies.

SIGNATURE: Joe Bel

DATE: 25/10/2019

## MARKER'S FEEDBACK:

(Continue overleaf if required)

MARKERS NAME:

GRADE/RESULT:





# ASSIGNMENT MARKET RESEARCH AND FORECASTING

## Question 1

The first step is looking the scatterplot matrix to see if there is any relationship among the variables. According to the graph (see Appendix 1 -A) there are some strong relationship among the variables. For example, *Ratio of males to females* and *proportion of divorced people* shows a strong negative linear relation or *Proportion of widowed people* and *Proportion of people aged 65-84* shows a very good positive linear relation. Therefore, a principal component analysis can be a suitable technique to reduce the dimensionality of this dataset.

- a) Although the variables seem arising on equal footing, *the variable Ratio of males to females* is a ratio and the others are all proportions. This fact is enough to lead to me to the decision of using the correlation matrix for the PCA (I do not bother to check the variability of variables because the diversity of the unit of measure is a sufficient criteria)
- b) Since correlation matrix is the decided method there are four criteria to check. The Scree plot (see Appendix 1 -B) shows the elbow at the 6<sup>th</sup> component, therefore according to this criterion the number of principal components is 5 (number of eigenvalues by last elbow less 1). The next two criteria are shown through the *total variance explained table* (see Appendix 1 -C). Notice that for the sake of finding the appropriate number of principal components the analysis was run with fixed number of factors equal to 8. The number of eigenvalues exceeding 1 is 3. By looking at the *proportion of variance explained*, three principal components account for 73,38 % of the total variation of the dataset. Since adding an extra component would not increase the total variation significantly (12% increase) I stick with 3 principal components. The last criterion is the parallel procedure (see Appendix 1 -D). This graph compares the real eigenvalues with the average eigenvalues from the simulation. The crossover point occurs at 2 principal components. Below the summary have been shown as well as the final decision:

Scree Plot: 5

Number of eigenvalues exceeding 1: 3

Proportion of variance explained: 3

Parallel procedure: 2

In conclusion, I decide to keep 3 principal components since 50% of the criteria found it as the right number of PCs.

- c) Here the equation of the two principal components are displayed:

$$PC1 = -0.718M1\_F1 - 0.397y15\_29\_pr - 0.217y30\_49 + 0.587y50\_64\_pr + 0.802y65\_84\_pr + 0.893W\_pr + 0.265D\_pr - 0.617Ec\_a\_pr$$

$$PC2 = -0.408M1\_F1 + 0.673y15\_29\_pr - 0.627y30\_49 + 0.041y50\_64\_pr - 0.246y65\_84\_pr + 0.011W\_pr + 0.693D\_pr + 0.295Ec\_a\_pr$$

The component matrix helps us with the interpretation of the principal components (see Appendix 1 -E). I am using a 0.650 cut-off (absolute value).

The first principal component contrasts *Proportion of people aged 65-84 (y65\_84\_pr)* and *Proportion of widowed people (W\_pr)* with *Ratio of males to females (M1\_F1)* and *Proportion of economically active population (Ec\_a\_pr)*.

The second principal component contrasts *Proportion of people aged 15-29 (y15\_29\_pr)* and *Proportion of divorced people (D\_pr)* with *Proportion of people aged 30-49 (y30\_49\_pr)*.

The PC1 might be interpreted as an indicator of population workforce, while the PC2 seems a bit hard to interpret. The eigenvectors are usually hard to interpret, however the main goal of PCA is reducing the dimensionality rather than give interpretation to the components.

- d) The scatterplot (see Appendix 1 -F) plots the two principal component scores together. The plot is what I expected because the data points are spread around without showing any relationship. In fact, the principal components are not correlated to each other. In addition, neither clusters nor outliers can be clearly identified (for outliers, no datapoint is greater or less than 3). Just for curiosity, nations such as Ireland and Iceland are “pushed” to the left side of the graph because those are the nations with a high ratio of males to female and large proportion of economically active population as well as small proportion of ‘old’ people and small proportion of widowed (see the PC1 interpretation).

## Question 2

The main objective of this question is whether factor analysis is a suitable technique for this dataset or not. There are four criteria we need to examine:

- We need to make sure that we have many correlations that exceeds 0.3 in absolute value. According to the correlation matrix (see Appendix 2 -A) 55/66 correlation exceeds 0.3 in absolute value. This means that more than 80% of the correlations are high. In general, correlations exceeding 0.30 provide evidence to indicate that there is enough commonality in the original variables to justify comprising factors.
- The next check is verifying whether the determinant is not zero or extremely close to zero. Unfortunately, the determinant is 2.599E-12 which is near to zero. Some books set a threshold of 0.00001 which means that our determinant is lower.

- Bartlett's test hypotheses are stated below:

Ho: Correlation matrix is an identity matrix

H1: Correlation matrix is not an identity matrix

We must ensure that the p-value is smaller than 0.05 and therefore reject the null hypothesis. Since the p-value is around 0 (see Appendix 2 -B), the test provided evidence that the observed correlation matrix is statistically different from a identity matrix, confirming that linear combination exists. In addition, the test is also telling us how close the data are to being close to a cigar shape rather than being in a spherical shape.

- KMO is high and can be classified as meritorious (see Appendix 2 -B). Although the value is high, this criterion might be improved because there is one variable *JPY* that have very small correlations with all the other observed variables. By deleting this variable, the KMO should increase.

In conclusion, a determinant smaller than the threshold suggests that the data is not appropriate for a factor analysis because of multicollinearity and the analysis cannot go further.

### Question 3

- a) Each graph will be commented below (see Appendix 3 -A).:

I) Raw data does not appear to come from a stationary process. In fact, there is a positive trend in the mean.

II) Transformed data fluctuates about a constant value and nearly all the variation is within a constant band. Therefore, transformed data appears to come from a stationary process. However, it can be considered a border line because it might be seen that there is a slight negative trend.

III) Transformed data does not appear to come from a stationary process. In fact, there is a negative trend in the mean.

IIII) Transformed data fluctuates about a constant value and all the variation is within a constant band. Therefore, transformed data appears to come from a stationary process.

- b) I decided to choose the simplest transformation of the data that appears to come from a stationary process (see Appendix 3 -B). Therefore, the first non-seasonal difference of the data will be used to show the ACF and PACF (see Appendix 3 -C).

It is important to know that we can fit an ARIMA model because the transformed data looks like come from a stationary process. According to the ACF and PACF, for the non-seasonal effect all the lags are inside the confidence limit suggesting no parameters.

For the seasonal effect lag 12,24 and 36 need to be evaluated. It seems like we have a decreasing ACF and a clear cut-off at lag 12 in a PACF. This would suggest a SAR(1). Therefore, I will try to fit the following model (see Appendix 3 -D):

ARIMA (0, 1, 0) (1, 0, 0)<sub>12</sub>

- c) After fitting that model, all the important model statistics and arima model parameters will be examined as well as residual diagnosis. The first try indicated that the constant term was not significant (see Appendix 3 -E) so in order to simplify the model the constant term was dropped (see Appendix 3 -F).

Next, the arima model without a constant term shows these features:

- All the coefficient in the model are significant.
- B-Ljung Q statistic do not reject the null hypothesis ( $\text{sig}=0.619$ ) which means that we are 95% confident that the autocorrelations in the residuals are white noise. Then, we conclude that this model is a candidate model and no more information is left in the residuals.
- Stationary R-squared (preferred to R-squared due to seasonality) is positive and tells us that the model is better than fitting a mean to the data.
- The model almost halved the RMSE (1191.832) comparing the initial standard deviation (2003.7907)

Note that the analysis could have done with other ARIMA models that might lead to better BIC (for example ARIMA (0, 1, 0) (1, 1, 0)<sub>12</sub>), but the goal of this point was to find an appropriate model rather than the best model.

ACF and PACF of the residuals indicates that there is no further pattern to remove. No information left in the residuals; in fact, all the spikes are within the significant limit (see Appendix 3 -G). All Box-Ljung values are not significant (see Appendix 3 -H).

To conclude the residuals diagnosis, we want to make sure the residuals are normally distributed (see Appendix 3 -I). The histogram looks like normally distributed, the Q-Q plot shows a linear trend and the Kolmogorov-Smirnov is not statistically significant which confirm the normality of the residuals.

## Question 4

As soon as we open the dataset, we first realize that there is no date variable. It is a good habit to set the units of the time variable first because it avoids having issues with some SPSS commands. Since the observation are recorded daily, I choose *cases are: Days* in SPSS (see Appendix 4 -A).

- a) The cross-correlation function (see Appendix 4 -B) enables us to see which previous days of rain influences the current observation of water level. We can see how the dependent variable is influenced by lagged independent variables. It is important to notice that we are only interested with non-positive lags.

By looking at the graph, *lag 2* is the only significant lag which means that the current water level of the Murray River is influenced by the daily rainfall 2 days before (I might have included *lag3* since it is very close to the confidence limit, however I prefer to stick with *lag2* only)

- b) The needed lags for the regression are *лаго*, *lag1*, *lag2*. We want to fit those predictor variables for the regression. Two new time series are created for this scope (see Appendix 4 -C).

After fitting the regression model to the data (see Appendix 4 -D) we first notice how small is the R squared values ( $R^2 = 0.028$ ). This is telling us that the relation between the variables is not linear at all. An indication of this poor relation can be examined by the scatterplot of the initial variables (Water Level with Rainfall). The scatterplot (see Appendix 4 -E) shows no relation at all between the two variables and this is mainly caused by the fact that most of the days did not rain (Rainfall = 0).

The  $DW_{OBS}$  is very low and this already might suggest that there is a strong positive autocorrelation in the residuals and therefore at lag 1. Those values below are the parameters that we need to test at 5% significant level (I have been using the values from the table provided in the class):

$$DW_{OBS} = 0.180$$

$$N = 363$$

$$k = 4$$

$$DW_L = 1.552$$

$$DW_U = 1.675$$

$H_0$ : residuals are not correlated

$H_1$ : residuals are correlated

As  $DW_{OBS} < DW_L$  we reject the null hypothesis and therefore state that there is correlation in the residuals

- c) Even though the residuals are correlated, they still should look like they come from a stationary process. However, according to the sequence chart (see Appendix 4 -F) the residuals does not look like coming from a stationary process (white noise) because they don't fluctuate around a mean and the variability seems increasing over time. This is a clear indication that the model is inadequate, and an ARIMA model cannot be fit.
- d) After transforming the dependent variable with the first order difference (see Appendix 4 -G) and fitting a new regression model on this new variable the sequence chart of the residual can be analysed.

The sequence chart of the residuals (see Appendix 4 -H) looks like coming from a stationary process now, and the residuals is a white noise. It is important to check that there is still a significant positive autocorrelation at lag 1 in the residuals since the "new" ( $DW_{OBS} = 1.399$ )  $< DW_L$ . Appendix 4 -I shows the model summary of the regression for the transformed dependent data.

In conclusion, since the residuals looks like comes from a stationary process, I can fit an arima to this regression model.

- e) We found out that the regression model is valid. By investigating the saved unstandardized residuals from the regression, an ARIMA model can be fit. From the ACF and PACF of the regression residuals (Appendix 4 -J) it is possible to note the significant autocorrelation at lag1 as expected from the DW test.

We could try to fit an ARIMA (2,0,3). Note that we are not interested in seeing if ACF or PACF decreases because we are dealing with residuals and therefore, we only focus on the significant spikes. In addition, the value 0 refers to the transformed dependent variable that I do not need to differentiate again.

An ARIMA model do not allow you to fit a model if some values are missing. Since *lag2 variable* do not have the first 2 observations, I removed the first 2 rows of the data to let the ARIMA works.

The important model statistics and arima model parameters will be examined as well as residual diagnosis. The first try indicated that some coefficients were not significant (see Appendix 4 -K) so in order to simplify the model those parameter terms were dropped one by one. It has been found that an appropriate model was an ARIMA (1,0,2) with only *lag2 variable* as predictor (see Appendix 3 -L).

Next, the ARIMA model with all the significant terms shows these features:

- All the coefficient in the model are significant.
- B-Ljung Q statistic do not reject the null hypothesis (sig=0.200) which means that we are 95% confident that the autocorrelations in the residuals are white noise. Then, we conclude that this model is a candidate model and no more information is left in the residuals.

ACF and PACF of the ARIMA residuals indicates that there is no further pattern to remove. No information left in the residuals; in fact, all the spikes are within the significant limit (see Appendix 4 -M). All Box-Ljung values are not significant (see Appendix 4 -N).

In conclusion, the residuals diagnoses are checked to make sure the residuals are normally distributed (see Appendix 4 -O). The histogram may look like normally distributed. However, the Q-Q plot does not show a linear trend and the Kolmogorov-Smirnov is statistically significant which means that the residuals are not normally distributed.

This is a border line case. According to otexts.com (one of the books provided by the lecturer) “A good forecasting method will yield residuals with the following properties:

1. The residuals are uncorrelated. If there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts.
2. The residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.

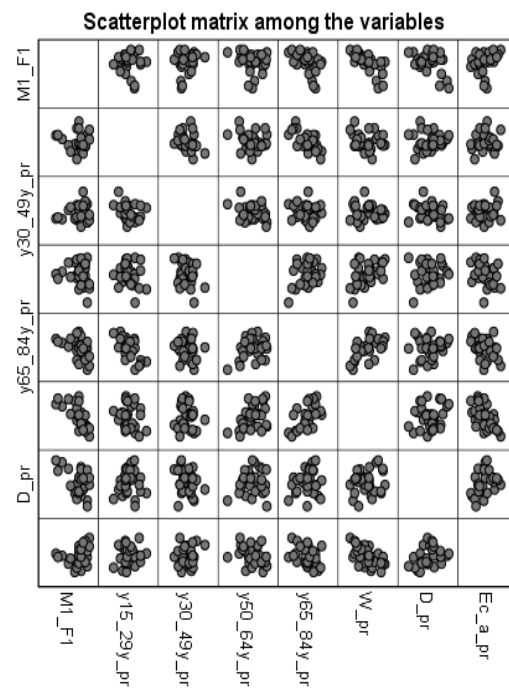
Our model has these essential properties. The article adds that checking the normality of the residuals is a useful but not necessary property.

In conclusion, the model is an appropriate (maybe not the best) model from the transformed data.

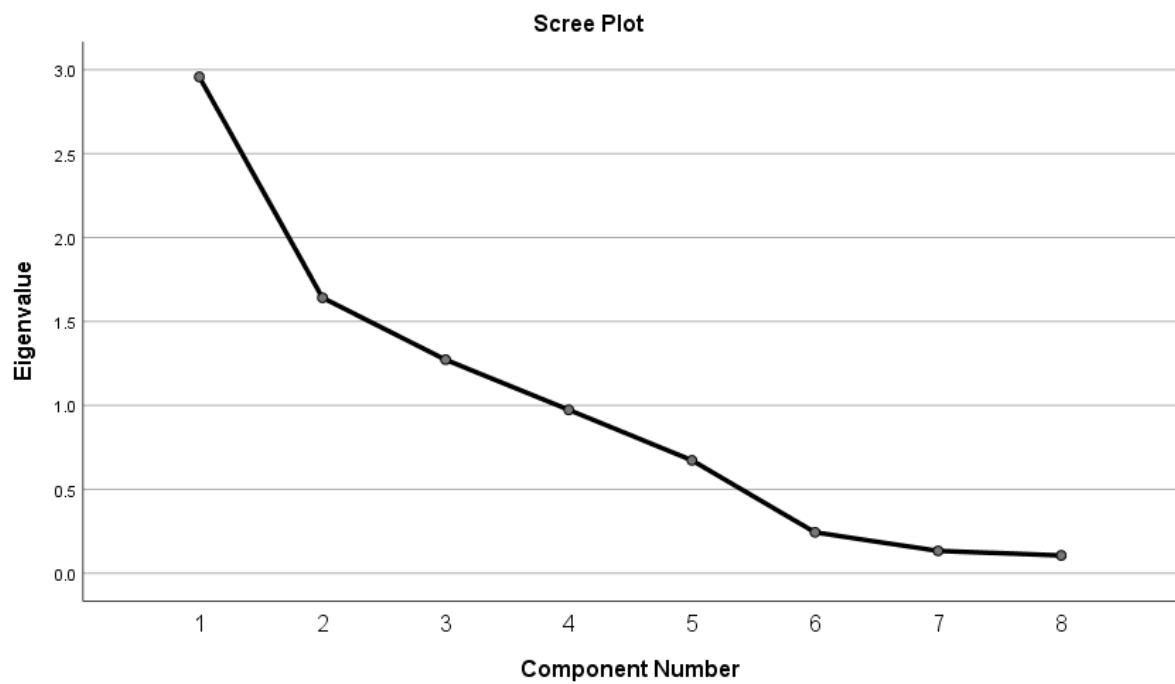
# APPENDIX

## appendix 1

### A) Scatterplot matrix



### B) Scree Plot





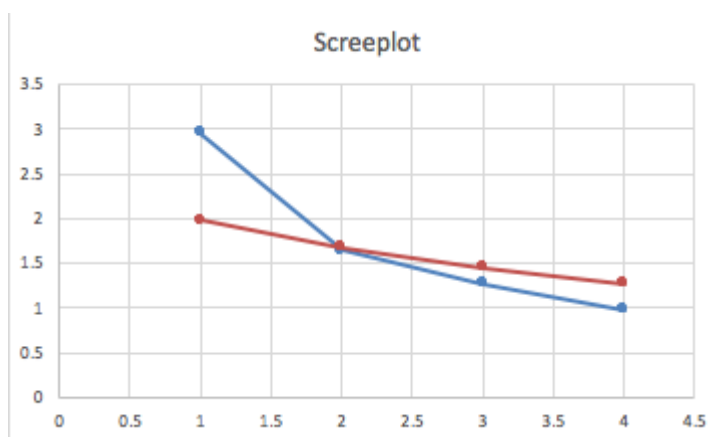
### C) Total variance explained

**Total Variance Explained**

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.957	36.961	36.961	2.957	36.961	36.961
2	1.641	20.518	57.479	1.641	20.518	57.479
3	1.272	15.902	73.381	1.272	15.902	73.381
4	.974	12.170	85.551	.974	12.170	85.551
5	.673	8.413	93.964	.673	8.413	93.964
6	.244	3.047	97.012	.244	3.047	97.012
7	.133	1.662	98.673	.133	1.662	98.673
8	.106	1.327	100.000	.106	1.327	100.000

Extraction Method: Principal Component Analysis.

### D) Parallel procedure

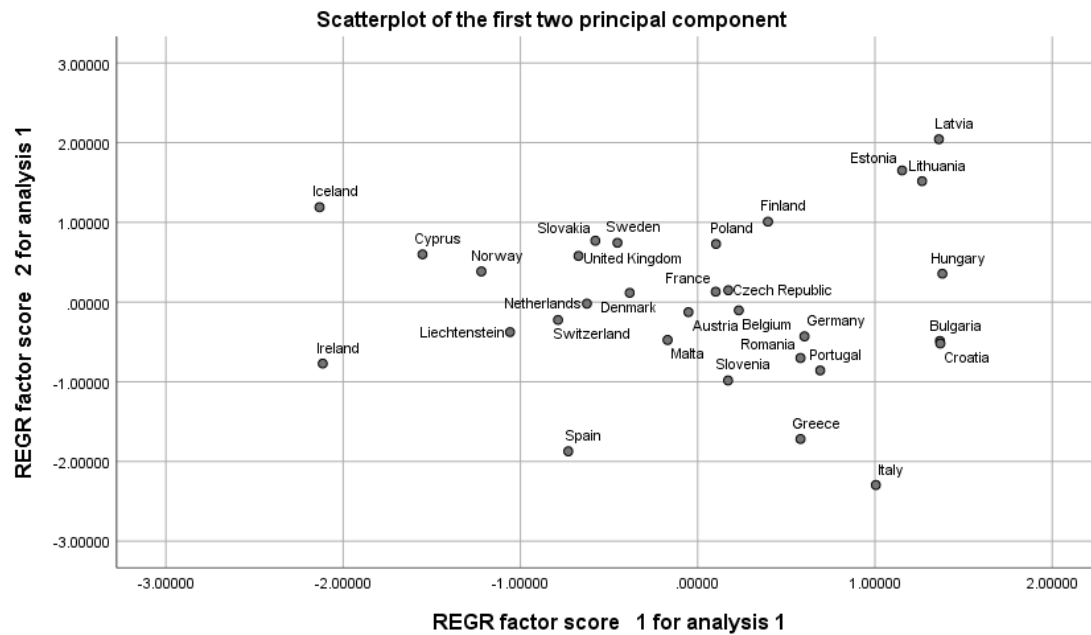


### E) Component matrix

**Component Matrix<sup>a</sup>**

	Component	
	1	2
M1_F1	-.718	-.408
y15_29y_pr	-.397	.673
y30_49y_pr	-.217	-.627
y50_64y_pr	.587	.041
y65_84y_pr	.802	-.246
W_pr	.893	.011
D_pr	.265	.693
Ec_a_pr	-.617	.295

## F) Scatter Plot graph



## appendix 2

### A) Correlation matrix

#### Factor Analysis

Correlation Matrix <sup>a</sup>													
	USD	CNY	JPY	EUR	KRW	GBP	SGD	THB	NZD	VND	CAD	CHF	
Correlation	USD	1.000	.963	.071	.795	.956	.786	.965	.960	.836	.941	.605	.896
	CNY	.963	1.000	-.108	.777	.959	.851	.982	.932	.898	.871	.573	.924
	JPY	.071	-.108	1.000	-.007	-.107	-.182	-.059	.035	-.291	.142	.088	-.087
	EUR	.795	.777	-.007	1.000	.827	.778	.787	.797	.686	.800	.686	.765
	KRW	.956	.959	-.107	.827	1.000	.797	.978	.936	.914	.861	.531	.932
	GBP	.786	.851	-.182	.778	.797	1.000	.798	.727	.711	.804	.687	.763
	SGD	.965	.982	-.059	.787	.978	.798	1.000	.945	.911	.848	.522	.951
	THB	.960	.932	.035	.797	.936	.727	.945	1.000	.802	.895	.595	.866
	NZD	.836	.898	-.291	.686	.914	.711	.911	.802	1.000	.696	.361	.897
	VND	.941	.871	.142	.800	.861	.804	.848	.895	.696	1.000	.761	.745
	CAD	.605	.573	.088	.686	.531	.687	.522	.595	.361	.761	1.000	.386
	CHF	.896	.924	-.087	.765	.932	.763	.951	.866	.897	.745	.386	1.000
a. Determinant = 2.599E-12													

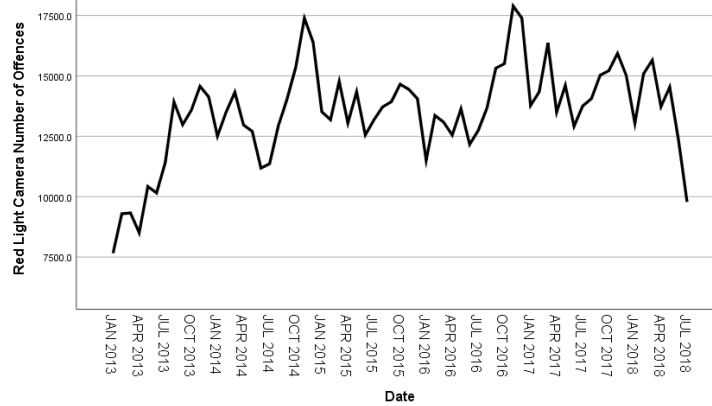
### B) KMO and Bartlett's test

#### KMO and Bartlett's Test

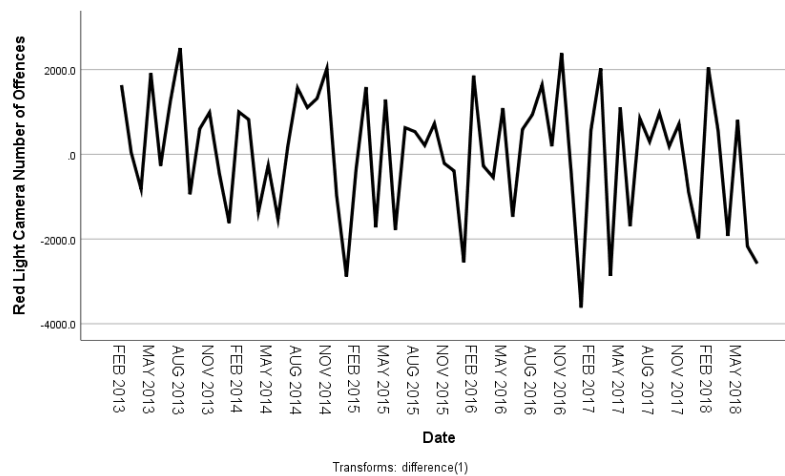
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.840
Bartlett's Test of Sphericity	Approx. Chi-Square	2618.681
	df	66
	Sig.	.000

## appendix 3

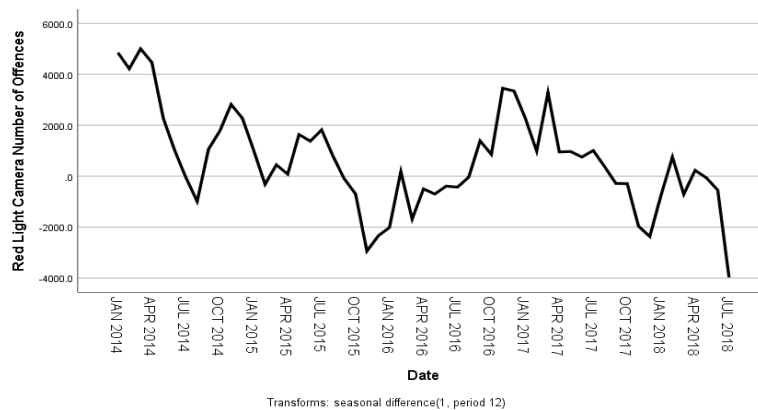
A) i) Sequence chart of the data



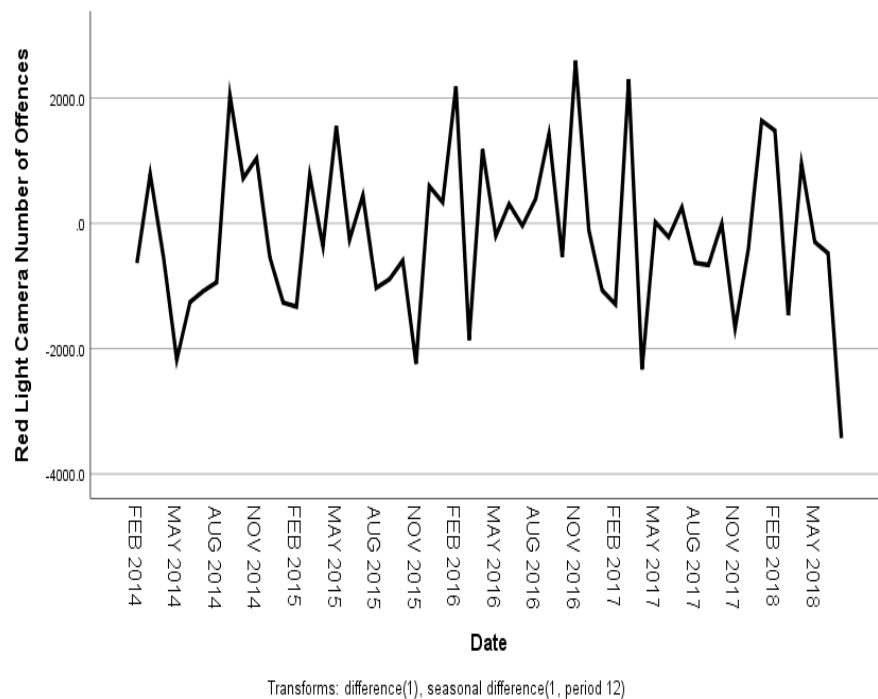
ii) Sequence chart of first non-seasonal difference of the data



iii) Sequence chart of first order-seasonal difference of the data



iv) Sequence chart of first non-seasonal difference and first order seasonal difference of the data



## B) Description on initial model

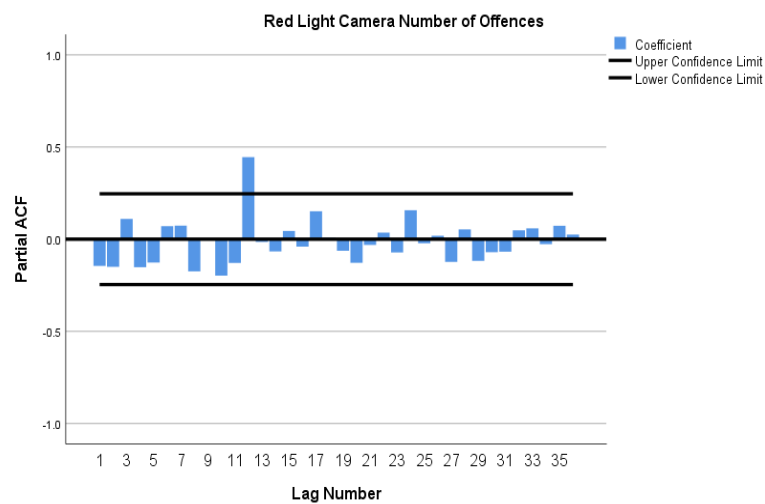
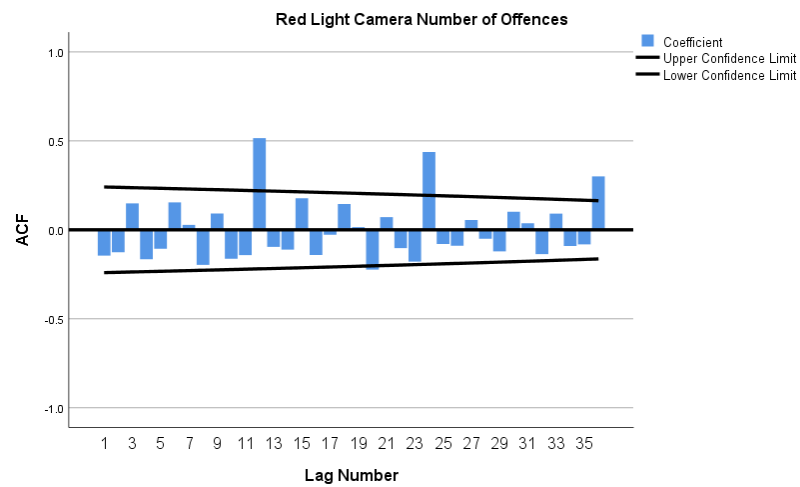
Model Description	
Model Name	MOD_16
Series Name	1
Transformation	None
Non-Seasonal Differencing	1
Seasonal Differencing	0
Length of Seasonal Period	12
Maximum Number of Lags	36
Process Assumed for Calculating the Standard Errors of the Autocorrelations	Independence(white noise) <sup>a</sup>
Display and Plot	All lags

Applying the model specifications from MOD\_16

a. Not applicable for calculating the standard errors of the partial autocorrelations.



### C) ACF AND PACF for up to 36 lags



### D) Model Description

#### Model Description

			Model Type
Model ID	Red Light Camera Number of Offences	Model_1	ARIMA (0,1,0) (1,0,0)

E) Model statistics and ARIMA parameter with constant.

Model Statistics													
Model	Number of Predictors	Stationary R-squared	R-squared	Model Fit statistics						Ljung-Box Q(18)			Number of Outliers
				RMSE	MAPE	MAE	MaxAPE	MaxAE	Normalized BIC	Statistics	DF	Sig.	
Red Light Camera Number of Offences-Model_1	0	.329	.600	1200.249	7.106	936.086	31.804	3112.000	14.308	14.788	17	.611	0

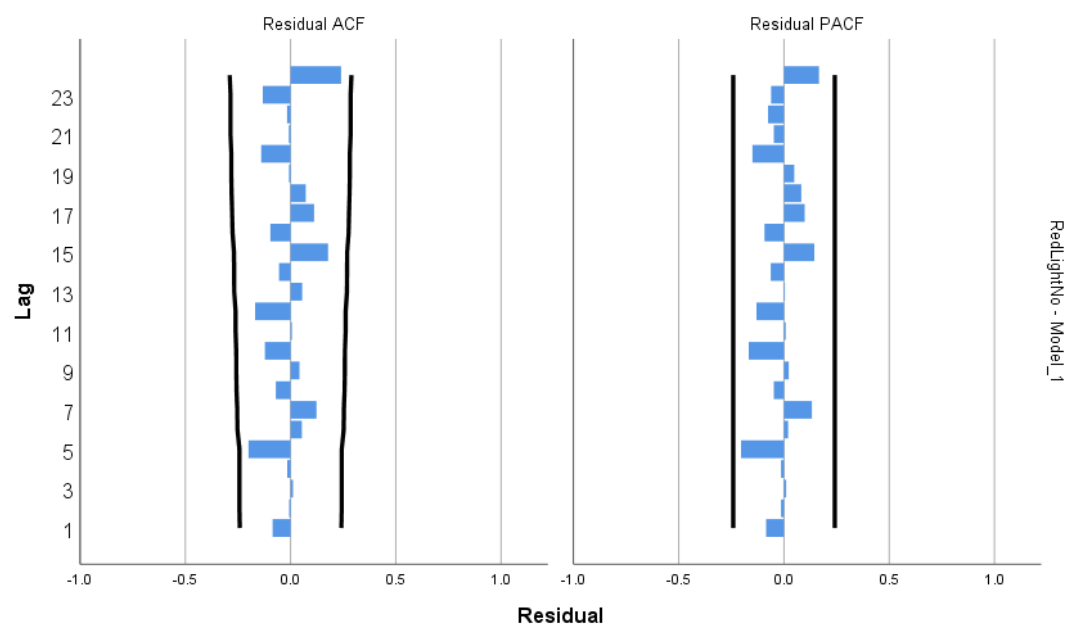
ARIMA Model Parameters							
Red Light Camera Number of Offences-Model_1	Red Light Camera Number of Offences	No Transformation					
			Constant	Estimate	SE	t	Sig.
			Difference	1			
			AR, Seasonal Lag 1	.613	.106	5.792	.000

F) Model statistics and ARIMA parameter without constant.

Model Statistics													
Model	Number of Predictors	Stationary R-squared	R-squared	Model Fit statistics						Ljung-Box Q(18)			Number of Outliers
				RMSE	MAPE	MAE	MaxAPE	MaxAE	Normalized BIC	Statistics	DF	Sig.	
Red Light Camera Number of Offences-Model_1	0	.328	.600	1191.832	7.100	935.459	31.676	3099.477	14.230	14.670	17	.619	0

ARIMA Model Parameters							
Red Light Camera Number of Offences-Model_1	Red Light Camera Number of Offences	No Transformation					
			Difference	Estimate	SE	t	Sig.
			AR, Seasonal Lag 1	.613	.104	5.889	.000
			Difference	1			

G) ACF and PACF of the residuals.



## H) Box- Ljung values

### Autocorrelations

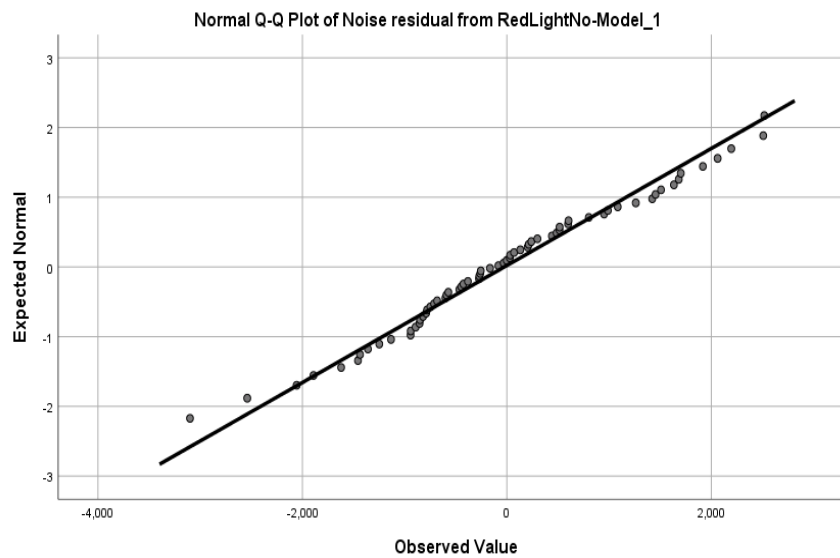
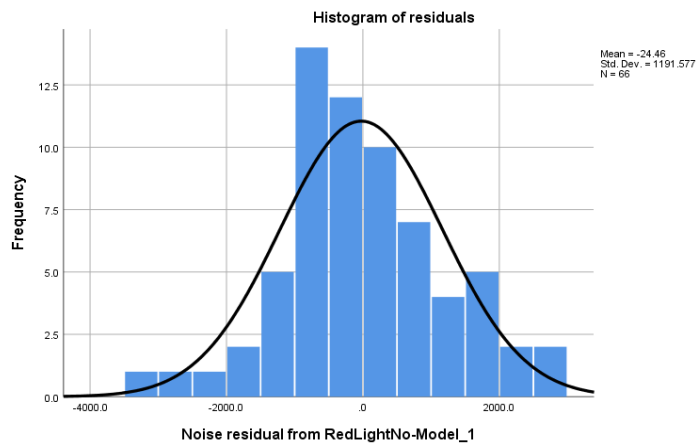
Series: Noise residual from RedLightNo-Model\_1

Lag	Autocorrelation	Std. Error <sup>a</sup>	Box-Ljung Statistic		
			Value	df	Sig. <sup>b</sup>
1	-.085	.120	.500	1	.479
2	-.007	.119	.503	2	.778
3	.012	.118	.513	3	.916
4	-.016	.118	.531	4	.970
5	-.200	.117	3.471	5	.628
6	.054	.116	3.689	6	.719
7	.123	.115	4.841	7	.679
8	-.070	.114	5.224	8	.733
9	.042	.113	5.365	9	.801
10	-.122	.112	6.563	10	.766
11	.006	.111	6.566	11	.833
12	-.168	.110	8.921	12	.710
13	.055	.109	9.176	13	.760
14	-.055	.108	9.432	14	.802
15	.178	.107	12.220	15	.662
16	-.096	.106	13.050	16	.669
17	.111	.104	14.185	17	.654
18	.072	.103	14.670	18	.684
19	-.007	.102	14.675	19	.743
20	-.140	.101	16.579	20	.680
21	-.008	.100	16.586	21	.736
22	-.017	.099	16.614	22	.784
23	-.132	.098	18.438	23	.733
24	.240	.097	24.598	24	.428

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

## I) Residuals Normality



### Tests of Normality

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Noise residual from RedLightNo-Model_1	.069	66	.200 <sup>*</sup>	.985	66	.611

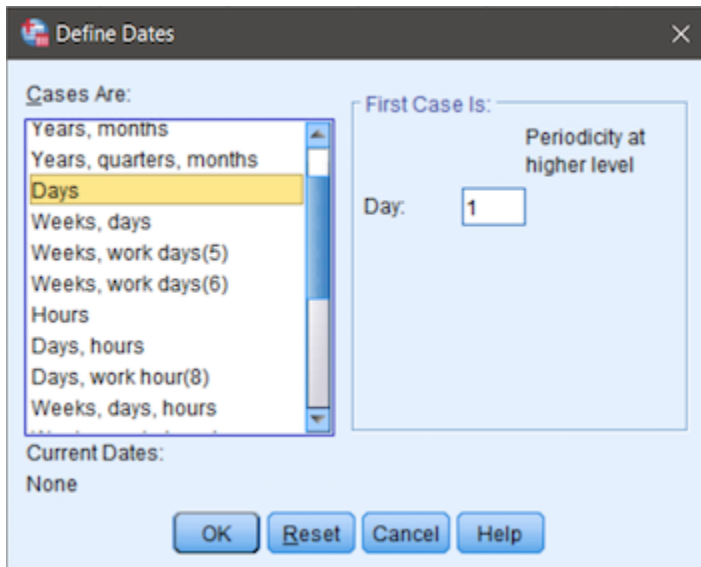
\*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



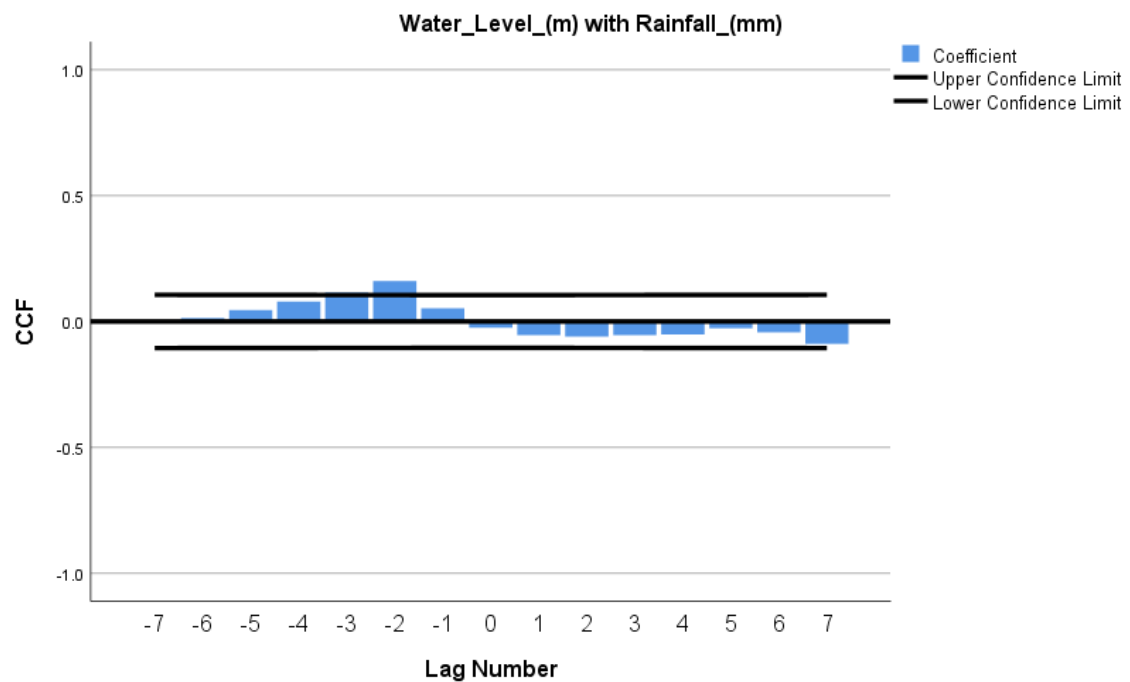
## appendix 4

### A) Define dates



The 'Define Dates' dialog box is shown. It has a title bar with a close button. The 'Cases Are:' section contains a list box with the following options: 'Years, months', 'Years, quarters, months', 'Days' (highlighted), 'Weeks, days', 'Weeks, work days(5)', 'Weeks, work days(6)', 'Hours', 'Days, hours', 'Days, work hour(8)', and 'Weeks, days, hours'. The 'First Case Is:' section has a 'Day:' label and a text box containing '1'. Below this is the text 'Periodicity at higher level'. The 'Current Dates:' section shows 'None'. At the bottom are four buttons: 'OK', 'Reset', 'Cancel', and 'Help'.

### B) CCF of water level with Rainfall



C) Creation of new two time series

Created Series					
	Series Name	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
		First	Last		
1	Rainfa_1	2	365	364	LAGS(Rainfall_mm,1)

Created Series					
	Series Name	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
		First	Last		
1	Rainfa_2	3	365	363	LAGS(Rainfa_1, 1)

D) Regression output

Model Summary <sup>b</sup>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.168 <sup>a</sup>	.028	.020	.306521	.180

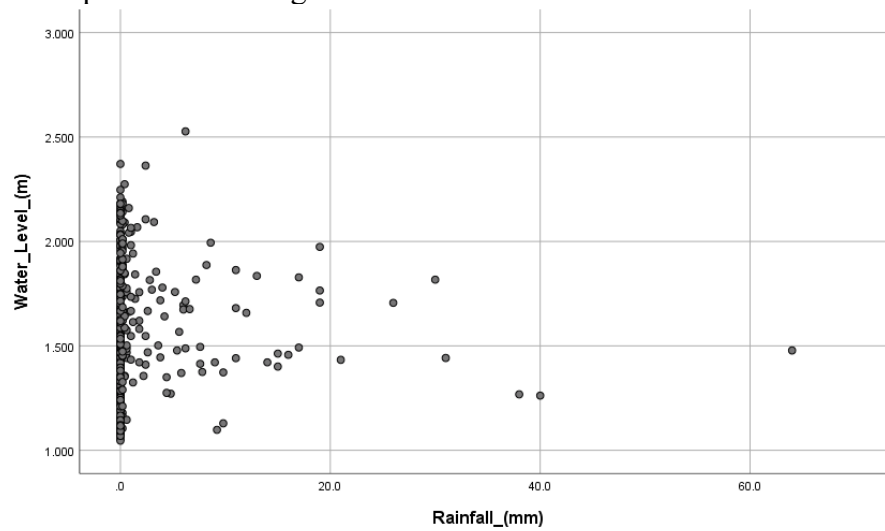
a. Predictors: (Constant), LAGS(Rainfa\_1,1), Rainfall\_(mm), LAGS(Rainfall\_mm,1)

b. Dependent Variable: Water\_Level\_(m)

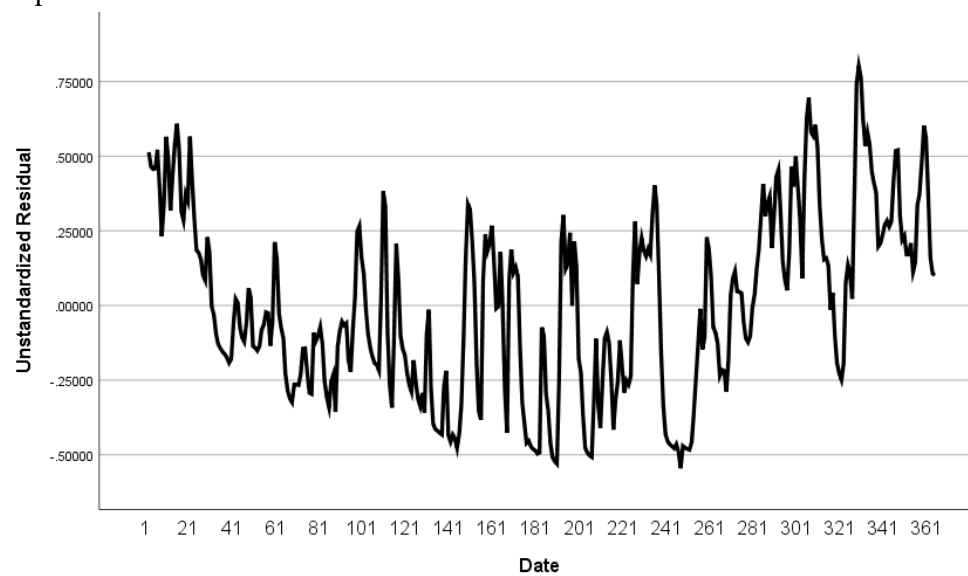
Coefficients <sup>a</sup>						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.576	.018		87.142	.000
	Rainfall_(mm)	-.001	.003	-.029	-.554	.580
	LAGS(Rainfall_mm,1)	.002	.003	.039	.737	.462
	LAGS(Rainfa_1,1)	.008	.003	.157	3.004	.003

a. Dependent Variable: Water\_Level\_(m)

E) Scatterplot between original variables



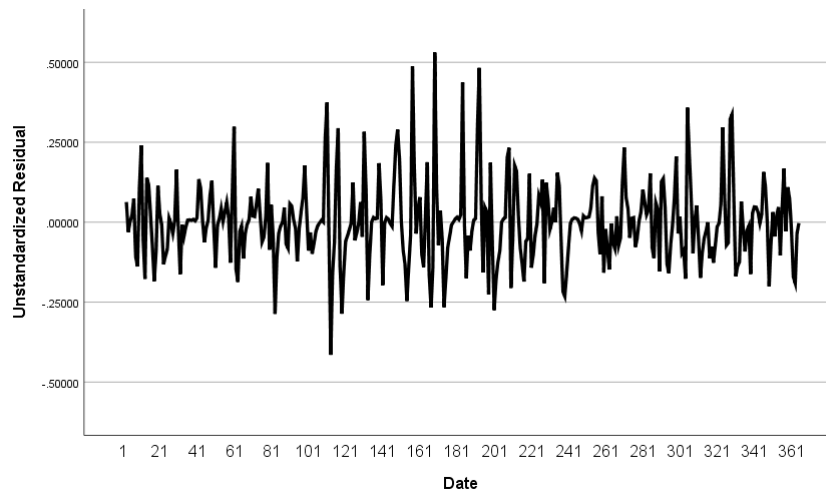
F) Sequence chart of the residuals



G) Creation of a new time series (Water level DIFF)

Created Series					
	Series Name	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
		First	Last		
1	Water__1	2	365	364	DIFF(Water_Level_m,1)

## H) Sequence chart of the residuals (Transformed data)



## I) Model summary of regression (Transformed data)

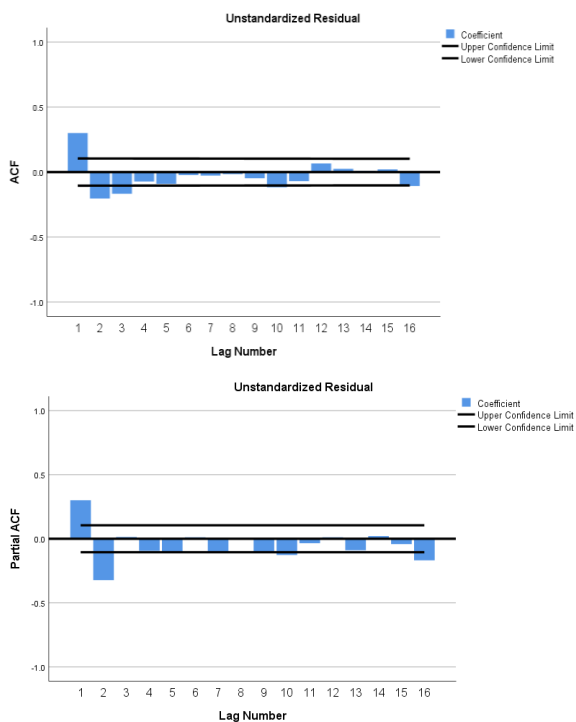
**Model Summary<sup>b</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.301 <sup>a</sup>	.091	.083	.127074	1.399

a. Predictors: (Constant), LAGS(Rainfa\_1,1), Rainfall\_(mm), LAGS(Rainfall\_mm,1)

b. Dependent Variable: DIFF(Water\_Level\_m,1)

## J) ACF and PACF of the Regression Residuals





K) Model statistics and ARIMA parameter (not significant parameters)

**Model Description**

				Model Type
Model ID	DIFF(Water_Level_m,1)	Model_1		ARIMA(2,0,3)

ARIMA Model Parameters										
				Estimate	SE	t	Sig.			
DIFF(Water_Level_m,1)- Model_1	DIFF(Water_Level_m,1)	No Transformation	Constant		-.013	.003	-4.210	.000		
			AR	Lag 1	-.021	.159	-.133	.894		
				Lag 2	.532	.138	3.843	.000		
			MA	Lag 1	-.395	.149	-2.651	.008		
				Lag 2	.788	.086	9.162	.000		
				Lag 3	.504	.075	6.692	.000		
			Rainfall_(mm)	No Transformation	Numerator	Lag 0	.000	.001	-.322	.747
			LAGS(Rainfall_mm,1)	No Transformation	Numerator	Lag 0	.001	.001	1.732	.084
	LAGS(Rainfa_1,1)	No Transformation	Numerator	Lag 0	.006	.001	6.160	.000		

L) Model statistics and ARIMA parameter (all significant parameters)

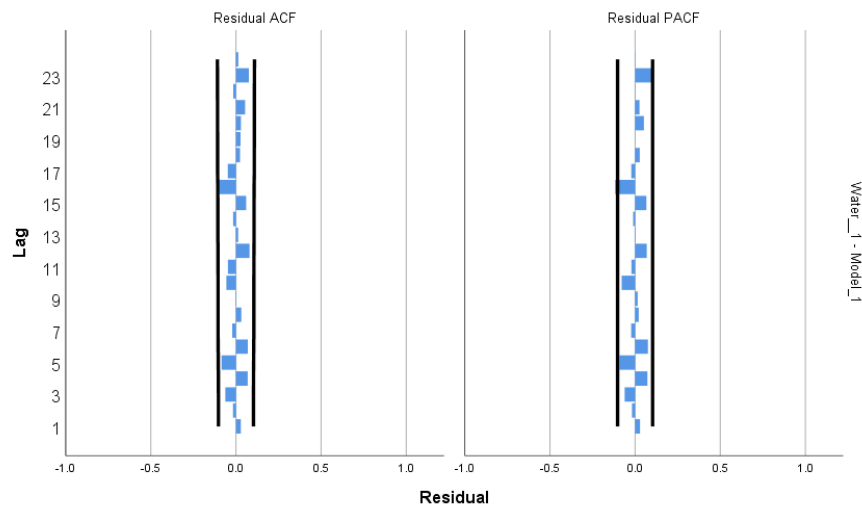
**Model Description**

				Model Type
Model ID	DIFF(Water_Level_m,1)	Model_1		ARIMA(1,0,2)

ARIMA Model Parameters								
				Estimate	SE	t	Sig.	
DIFF(Water_Level_m,1)- Model_1	DIFF(Water_Level_m,1)	No Transformation	Constant		-.010	.002	-4.885	.000
			AR	Lag 1	.753	.048	15.812	.000
			MA	Lag 1	.441	.055	8.082	.000
				Lag 2	.515	.048	10.793	.000
			LAGS(Rainfa_1,1)	No Transformation	Numerator	Lag 0	.005	.001

Model Statistics													
Model	Number of Predictors	Stationary R-squared	R-squared	RMSE	Model Fit statistics					Ljung-Box Q(18)			Number of Outliers
					MAPE	MAE	MaxAPE	MaxAE	Normalized BIC	Statistics	DF	Sig.	
DIFF(Water_Level_m,1)-Model_1	1	.290	.290	.112	208.872	.082	8670.909	.528	-4.289	19.321	15	.200	0

### M) ACF and PACF of the ARIMA Residuals



### N) Box-Ljung values

#### Autocorrelations

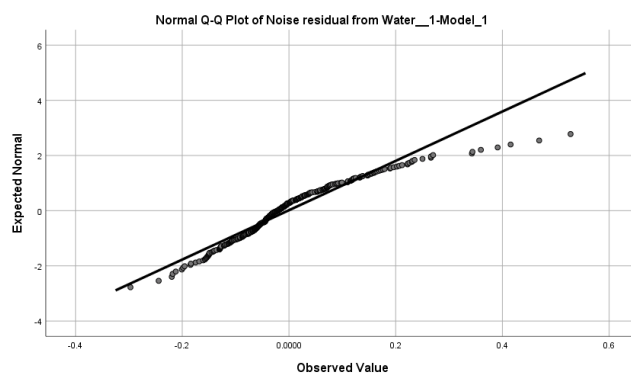
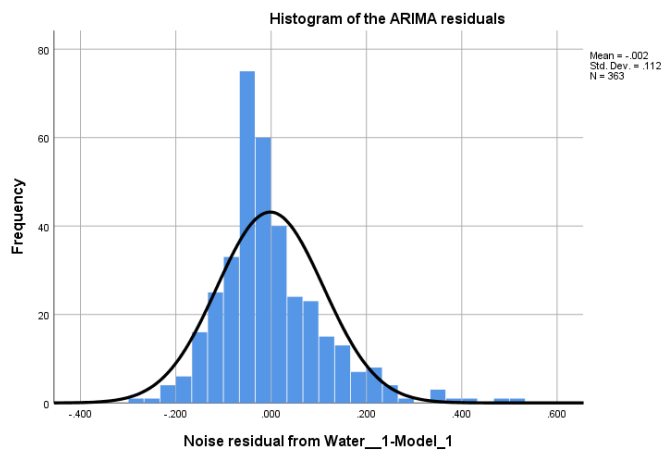
Series: Noise residual from Water\_\_1-Model\_1

Lag	Autocorrelation	Std. Error <sup>a</sup>	Box-Ljung Statistic		
			Value	df	Sig. <sup>b</sup>
1	.029	.052	.303	1	.582
2	-.018	.052	.417	2	.812
3	-.062	.052	1.839	3	.606
4	.069	.052	3.595	4	.464
5	-.084	.052	6.230	5	.284
6	.069	.052	8.014	6	.237
7	-.022	.052	8.193	7	.316
8	.032	.052	8.564	8	.380
9	-.004	.052	8.569	9	.478
10	-.056	.052	9.766	10	.461
11	-.047	.052	10.607	11	.477
12	.080	.051	13.002	12	.369
13	.013	.051	13.067	13	.443
14	-.017	.051	13.171	14	.513
15	.060	.051	14.536	15	.485
16	-.098	.051	18.220	16	.311

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

## O) Residuals Normality



### Tests of Normality

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Noise residual from Water__1-Model_1	.113	363	.000	.931	363	.000

a. Lilliefors Significance Correction

## REFERENCES

Hyndman, R.J and Athanasopoulos, G. (2018) Forecasting: Principles and Practice. 2<sup>nd</sup> Edition: O Texts. <https://otexts.com/fpp2/residuals.html>