Statistical Analysis Report

Author: Aurora

Background

Precipitation is an important factor for environment analysis. This report analysis the precipitation of Welland and Waterloo using time series in order that the precipitation of these two locations could be predicted in few year later.

Data Source

There are two datasets. One is about Wellan precipitation. This contains total precipitation by month from January 1995 to October 2004. The other is about Waterloo precipitation which contains total precipitation for the years 1950 to 2005.

Section 1:

Data Transformation and Cleaning (Description)

There are two columns in the Wellan dataset. Only the column of Precip is chosen and transformed to time series datatype.

Descriptive Data Analysis

Summary

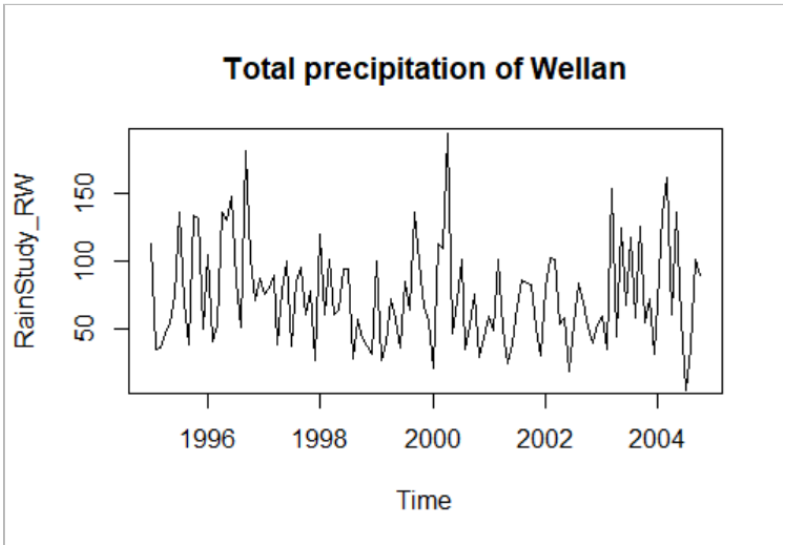
```
Min. 1st Qu.  Median    Mean  3rd Qu.    Max.
4.40   47.27   66.80   74.08   99.67   193.70
```

Descriptive Data

```
                     x
nbr.val      118.0000000
nbr.null       0.0000000
nbr.na         0.0000000
min            4.4000000
max          193.7000000
range        189.3000000
sum         8741.1000000
median        66.8000000
mean          74.0771186
SE.mean        3.3946275
CI.mean.0.95   6.7228820
var         1359.7725489
std.dev       36.8750939
coef.var       0.4977933
```

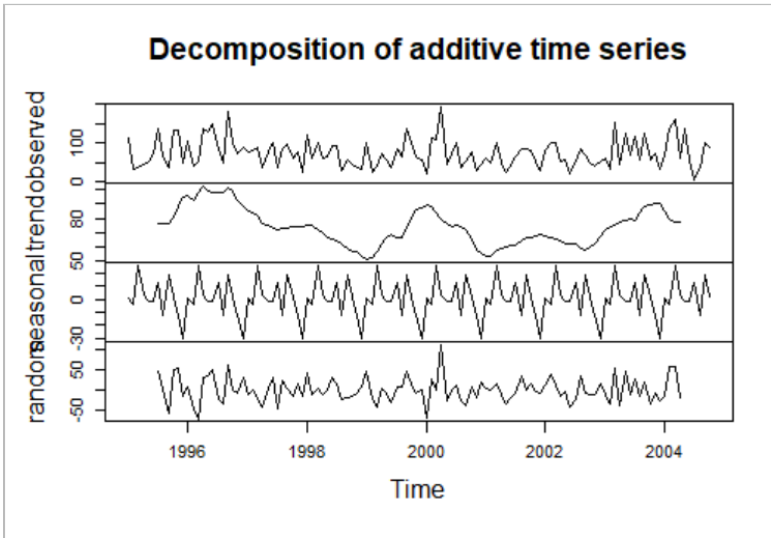
The result of summary data and descriptive data shows the data looks reasonable.

Plot time series



The graph shows the maximum value are all around year of 2000 and the minimum value is between 2004 and 2005. The average value between 1998 and 2003 seems to be lower than the other period. The value between 2003 and 2004 wave strongly.

Decompose time series data



The result of decomposition shows the data could be made up by three elements, seasonal, trend and random. Seasonal is a kind of periodic data and has a cycle of one year. It appears that the precipitation is higher in summer then drop down to the bottom in winter. Trend is a curve that wave gently. That is a moving average. It is sliding down from the top value in 1996 to the bottom in 1999, and then wave to climb up again to reach the second value in 2004.

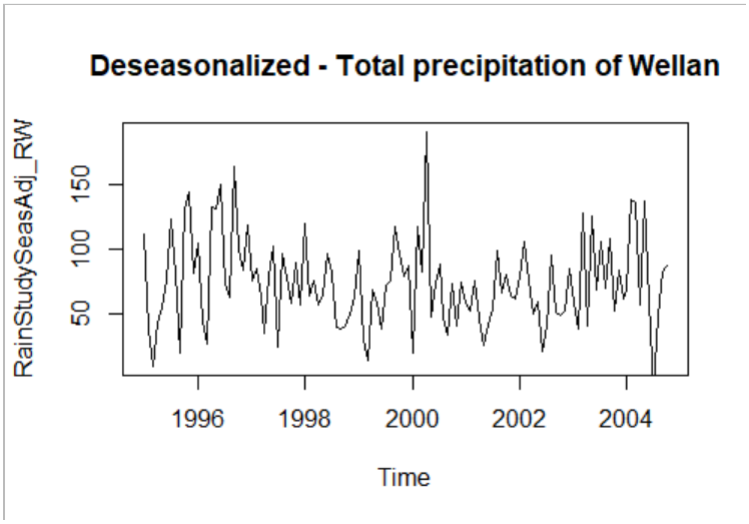
Check stationary

```
Augmented Dickey-Fuller Test

data:  RainStudy_RW
Dickey-Fuller = -4.7912, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

The result shows that p value is smaller that 0.05, which means the data is stationary. So the mean of every cycles are the same. And the variance of every cycle is the same.

Deseasonalize the precipitation



Deseasonalize the data is for getting rid of the impact of seasonal so that it could be easier to get the trend of data and the feature of the data. From the graph , without the seasonal , the curve is not so sharp. Especially, the curve between 1998 and 1999 is smoother so that the trend of this period is the value is going down first and then growing up.

Section 2:

Data Transformation and Cleaning (Description)

There are two columns in the Waterloo dataset. Only the column of Precip is chosen and transformed to time series datatype.

Descriptive Data Analysis

Summary

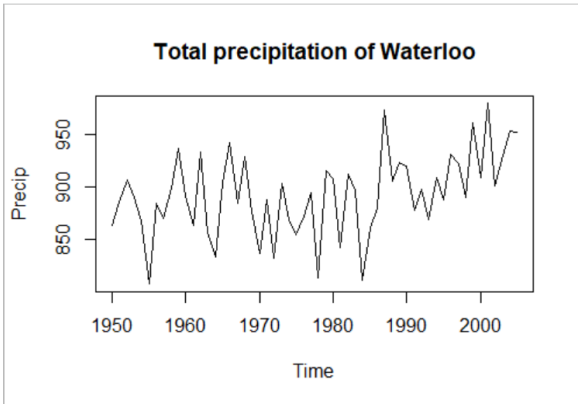
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
4.40   47.27   66.80   74.08   99.67  193.70
```

Descriptive Data

	Precip
nbr.val	56.00000000
nbr.null	0.00000000
nbr.na	0.00000000
min	807.79000000
max	979.87000000
range	172.08000000
sum	50027.55000000
median	892.76500000
mean	893.34910714
SE.mean	5.19859348
CI.mean.0.95	10.41821414
var	1513.42095373
std.dev	38.90271139
coef.var	0.04354704

The result of summary data and descriptive data shows the data looks reasonable

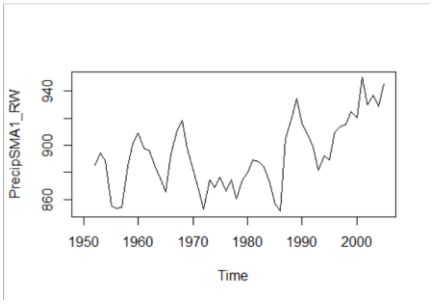
Plot time series



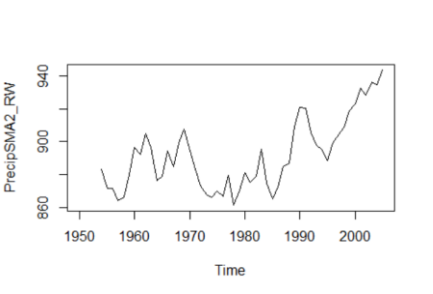
This graph shows the variance of precipitation value is big before year of 1990, and after that it is becoming smaller. And the mean before seems lower that the mean after 1990. And the precipitation reach to top value in 2000.

Simple Moving Average

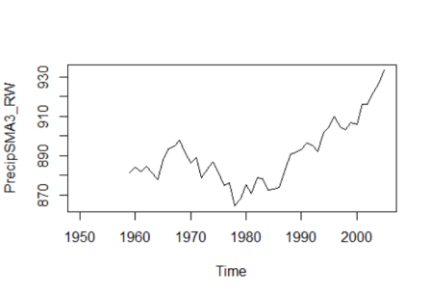
N=3



N=5



N=10



Compare with these three graphs, the result of N=5 is better, which use the values of previous five years to calculate. Because the third one(N=10) loss some detail information. The first one(N=3) waves too much, which means the trend is not so clear.

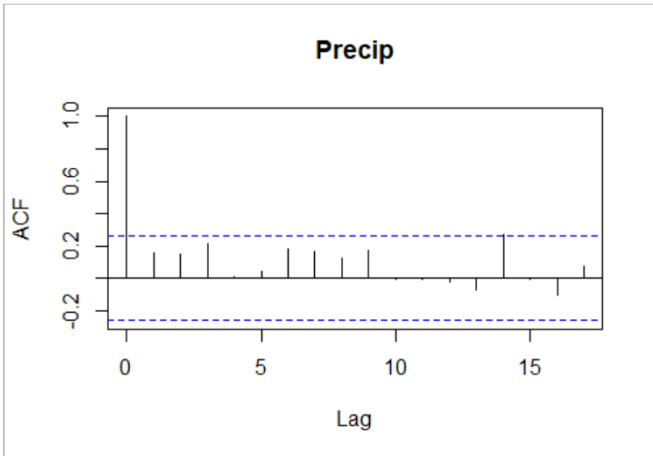
Check stationary

```
Augmented Dickey-Fuller Test

data:  Precip_RW
Dickey-Fuller = -3.5318, Lag order = 3, p-value = 0.04691
alternative hypothesis: stationary
```

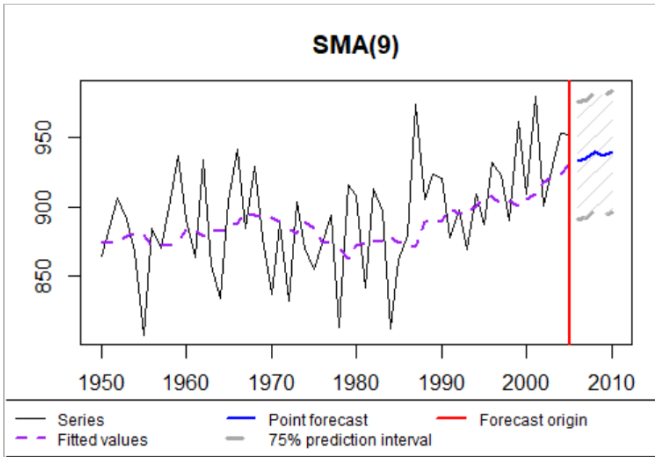
From the result, although the p value is 0.04691 which is smaller than 0.05, it is too close to 0.05. So, the data is stationary but not highly significant.

Autocorrelations



From the graph, except the lag0 which the precipitation itself, there is only lag14 that has positive correlation with lag0, but not so much, because it is just reach the value of blue line. So that means previous values seems not influence the current values.

Simple Moving Average Forecast



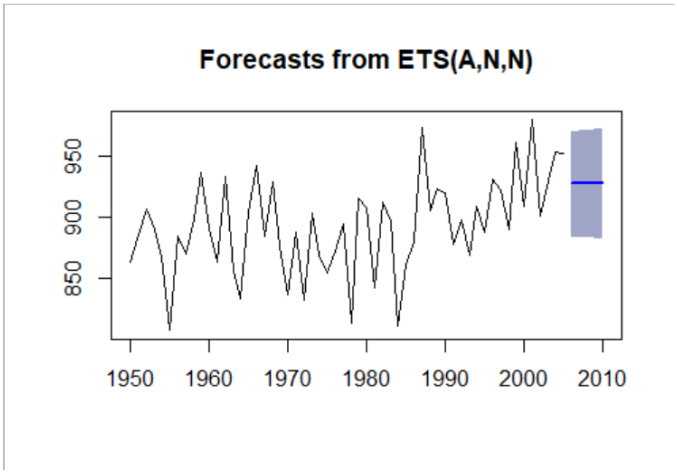

```
Time elapsed: 0.04 seconds
Model estimated: SMA(9)
Initial values were produced using backcasting.

Loss function type: MSE; Loss function value: 1282.8495
Error standard deviation: 36.4741
Sample size: 56
Number of estimated parameters: 2
Number of degrees of freedom: 54
Information criteria:
     AIC     AICc      BIC     BICc
563.7041 563.9305 567.7548 568.2105


Time Series:
Start = 2006
End = 2010
Frequency = 1
        Point forecast Lower bound (12.5%) Upper bound (87.5%)
2006       933.3911            890.9768            975.8054
2007       934.5435            891.8682            977.2187
2008       939.4794            896.4841            982.4747
2009       937.0649            893.6777            980.4521
2010       940.1754            896.3093            984.0416
```

The output of the model shows the loss value is 1281.8495. The AIC is 563.7041. The blue line in the graph forecasts the values of next five year. There is 75% probability of true value being in the shadow area.

Exponential Smoothing Forecast




```
Time elapsed: 0.09 seconds
Model estimated: ETS(ANN)
Persistence vector g:
 alpha
0.1428
Initial values were optimised.

Loss function type: MSE; Loss function value: 1349.906
Error standard deviation: 37.7666
Sample size: 56
Number of estimated parameters: 3
Number of degrees of freedom: 53
Information criteria:
     AIC     AICc      BIC     BICc
568.5574 569.0189 574.6334 575.5623


      Point Forecast    Lo 75    Hi 75
2006         927.752 884.7114 970.7927
2007         927.752 884.2736 971.2304
2008         927.752 883.8402 971.6639
2009         927.752 883.4110 972.0930
2010         927.752 882.9860 972.5181
```

The output of the model shows the loss value is 1349.906. The AIC is 568.5574. The blue line in the graph forecasts the values of next five year. There is 75% probability of true value being in the shadow area.

Compare the two forecasts

Compare with the two forecasts, the SMA is better than ETS, for the AIC of SMA is smaller than the AIC of ETS. And the loss value of SMA is smaller than the loss value of ETS. Also, the predictions of SMA is more sensible than the predictions of ETS.