Carnegie
Mellon
University
Africa

DATA SCIENCE

Spatial and Temporal Analysis of Rainfall and Vegetation Dynamics in Rwanda

Report

By: Agnes Nakalembe

Imported Libraries/modules

- tabulate
- pandas
- matplotlib.pyplot
- warnings
- matplotlib.ticker.MaxNLocator
- haversine.haversine , Unit
- scipy.optimize.curve_fit
- numpy
- sklearn.metrics.r2_score, mean_squared_error
- sklearn.model_selection.cross_val_score, KFold
- sklearn.base.BaseEstimator, RegressorMixin

Data Preparation and Initial Loading for District-Level Rainfall and Vegetation Analysis

Procedure

- The rainfall and vegetation data for the 30 districts in Rwanda was loaded into two separate data frames using pandas.
- The data frames were checked for missing values and the found missing values were then dropped.

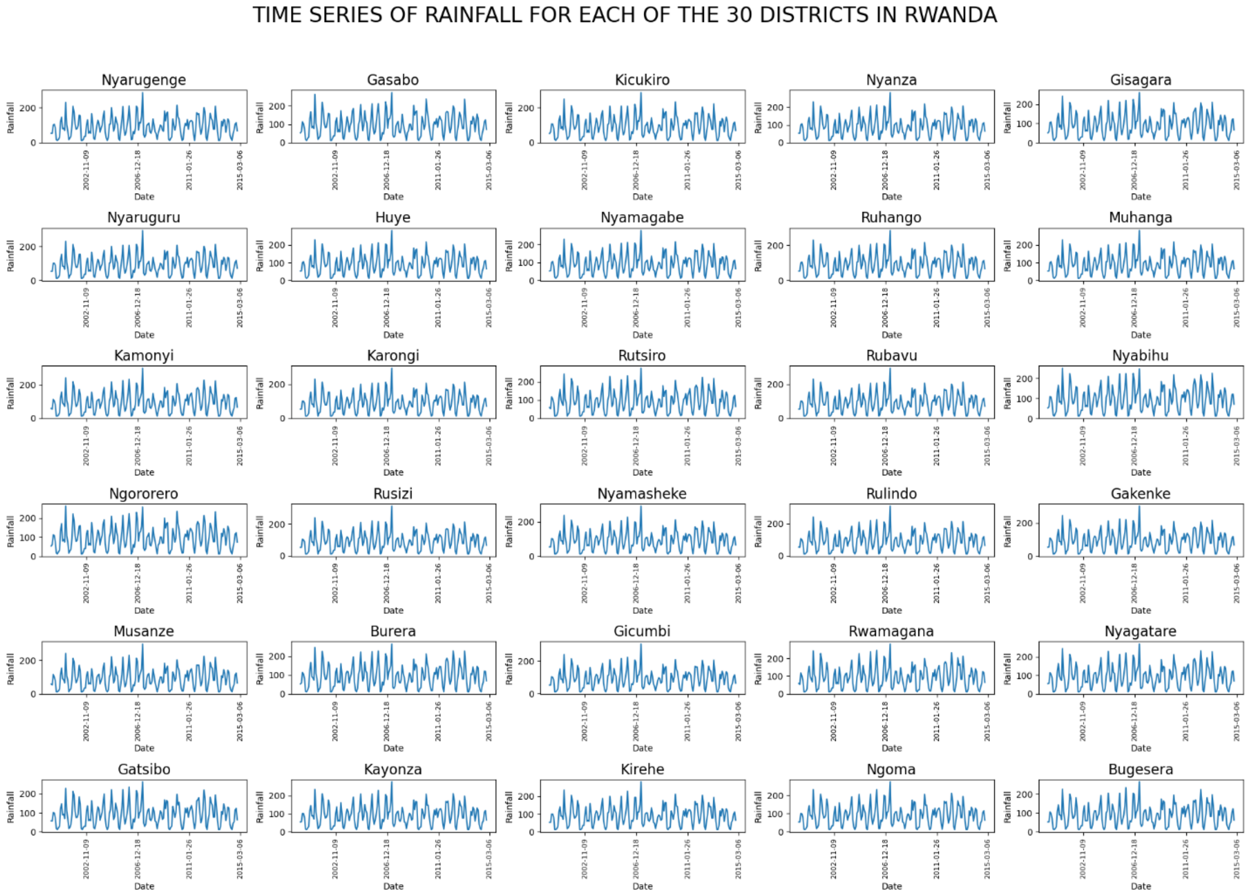
Time Series Visualization of Rainfall and Vegetation Indices Across Rwandan Districts

Procedure

- Firstly, a ‘Date’ column was created in both the rainfall and vegetation datasets using the provided Month column.
- Then using the rainfall data frame and the plot function in matplotlib, subplots of the time series for each of the 30 districts in Rwanda were done with the rainfall values on the y-axis and the date values on the x-axis.
- Subplots were also done using the vegetation index, again for each of Rwanda’s 30 districts.

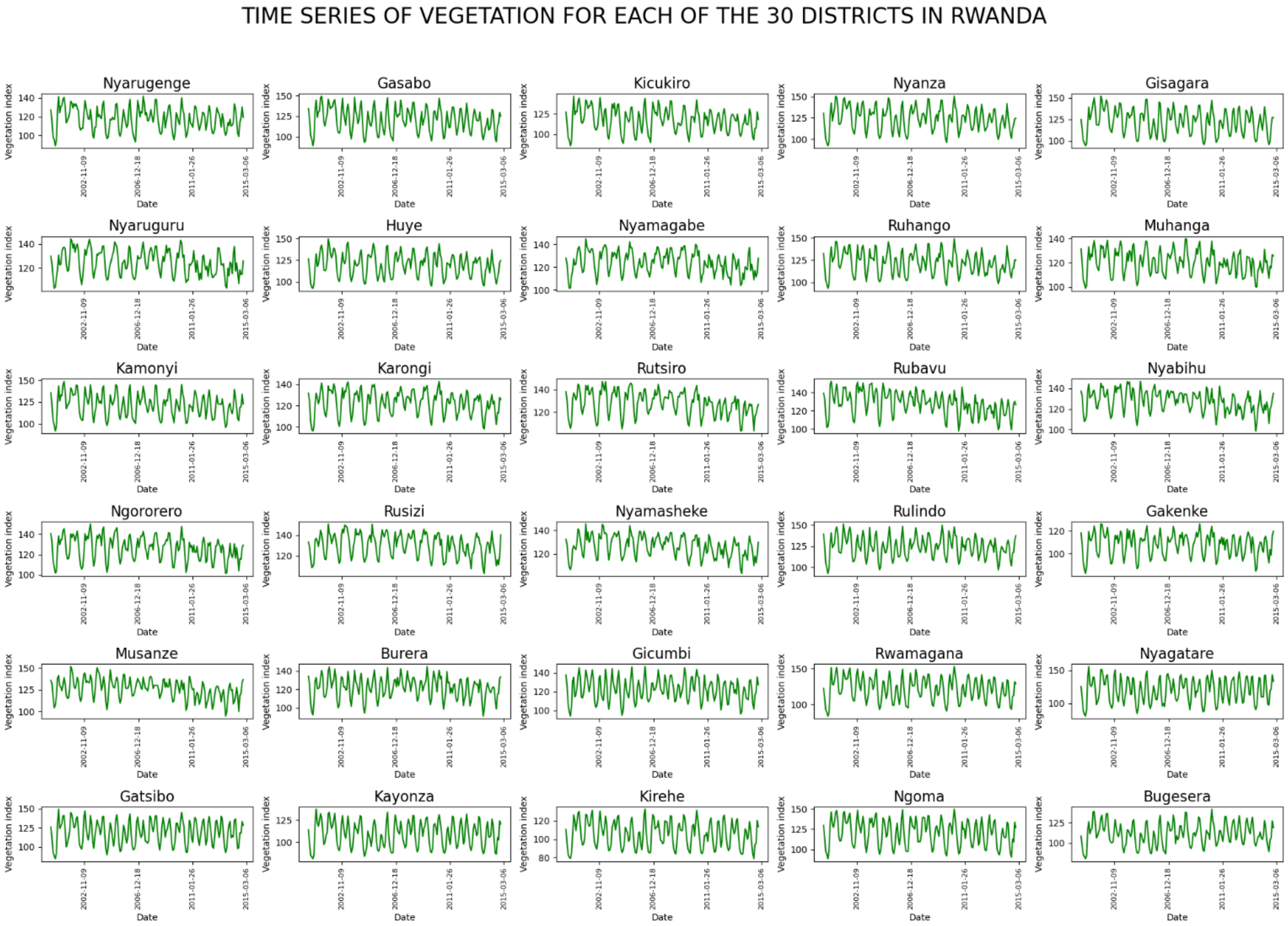
Results

Rainfall



Insights

- Approximately, all the 30 districts had the same trend of rainfall for the entire period.
- There is steady high and low peaks of rainfall in all the 30 districts. The high peaks are most probably experienced during the rainy season and the low peaks, during the dry season.



Insights

- There is fluctuations between high and low vegetation index for all the 30 districts. However, these fluctuations are somehow more rapid for some districts such as Gatsibo, Nyagatare, and less rapid for districts such as Rutsiro, Rusizi. These fluctuations can be attributed to the change of seasons between wet/rainy seasons and dry seasons that are experienced in Rwanda.
- In most districts, the vegetation index remains relatively higher than the rainfall lows, suggesting that vegetation remains intact during dry seasons. The vegetation's health declines but does not wither away entirely.

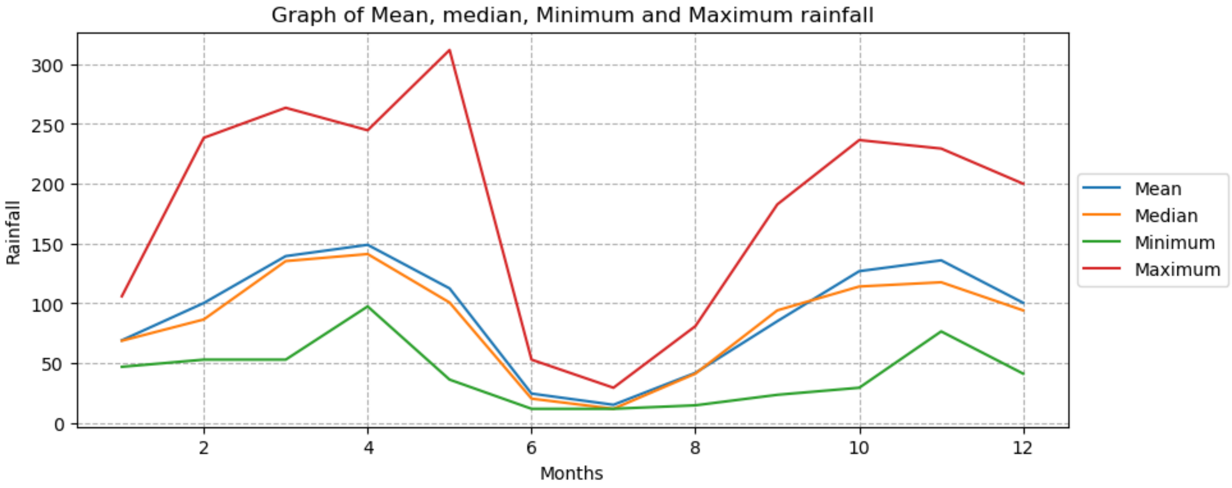
Statistical Analysis of Monthly Rainfall and Vegetation Indices

Procedure

- Having retrieved the rainfall data in all the 30 districts for each of the 12 months of the year, the monthly mean rainfall values were calculated using the pandas mean function, the minimum value using the min function, the median value using the median function and the maximum value using the max function.
- The above procedure was repeated using vegetation index data.
- Plots of these statistics against the months were then made using matplotlib's pyplot function.

Results

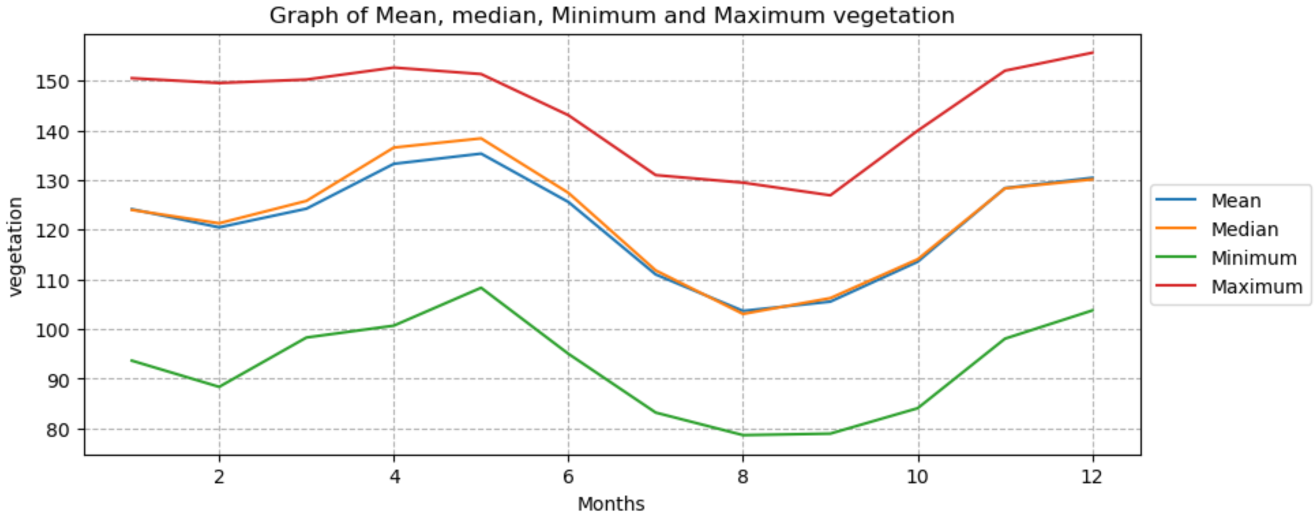
Rainfall



Insights

- By looking at the mean, median, and minimum rainfall curves, it is observed that these values are high between March and May. This agrees with the fact that Rwanda experiences a rainy season between march and May. The maximum rainfall occurred in May.
- From June to mid-September, Rwanda experiences lower rainfall, aligning with its long dry season. The dry season seems to be greatest in July, evidenced by the lowest rainfall values being in this month as per the plot.
- Rainfall values increase between October to November. This is because of the short rainy season that occurs in Rwanda during these months.
- Rainfall values start to decrease from December and continue to be low until February. This is because here is a dry season during this period in Rwanda. [1]

Vegetation



Insights

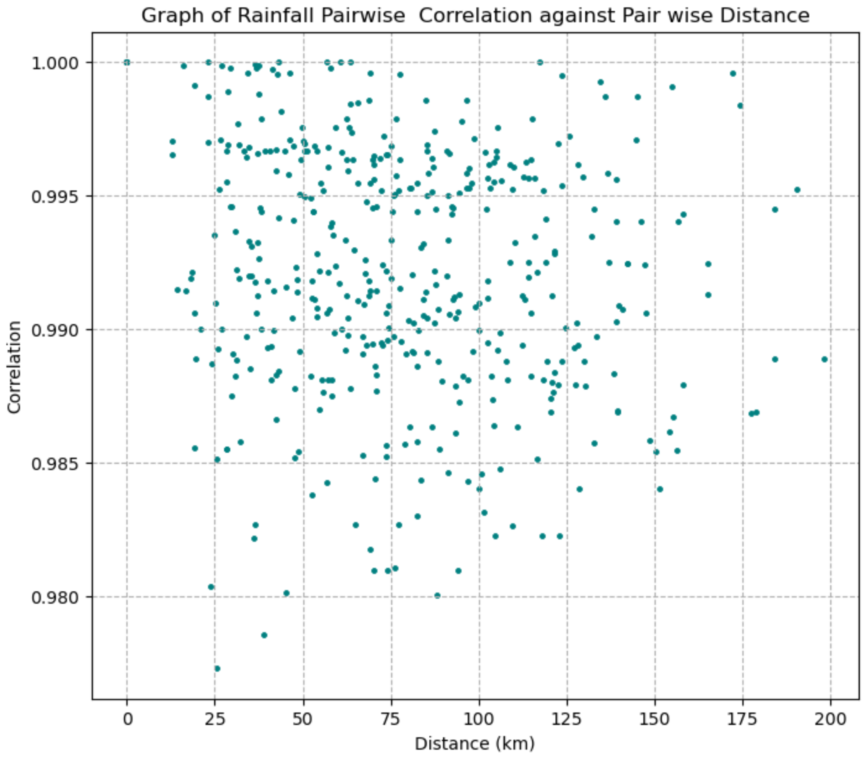
- Looking at the mean, median, and minimum vegetation index values, it is evident that the vegetation index has its peak values in May. This can be attributed to the fact that the vegetation benefits from the previous month's rainfall. As was noted from the rainfall plot earlier, the peaks of the rainfall occurred in April, hence the peak of vegetation being in May.
- Also, the lowest vegetation index values are in August. Once again, the fact that vegetation benefits from the rainfall of the previous month, is exposed, as it was observed that Rwanda experiences its lowest rainfall in July.

Spatial Correlation Analysis of Rainfall Across Districts

Procedure

- The rainfall correlations among districts were obtained using pandas corr function on the rainfall data.
- Using the haversine function in the haversine package, the distance between districts were computed, using the Rwanda district centroid-latitude-longitude dataset, with the Unit value set to kilometers. This function makes use of the latitude and longitude values of two districts to calculate the distance between these districts.
- A scatter plot of correlation against distance was then made using pyplot's scatter function.
- A model was fitted using the correlation and distance values, and a new graph was made that includes this fitted model.

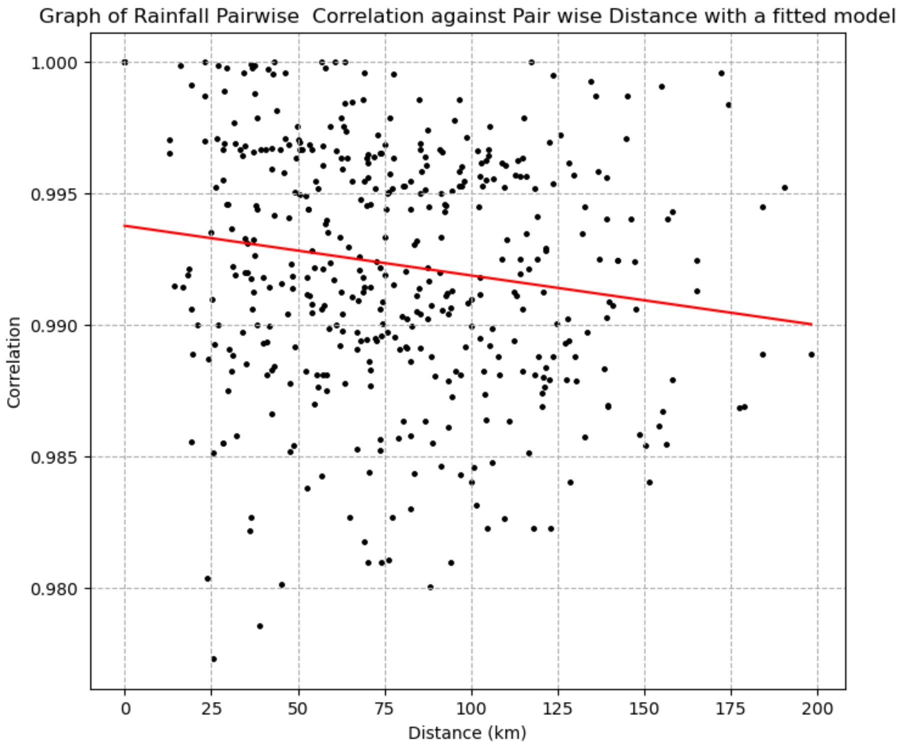
Results



Insights

- Overall, the correlation between districts' rainfall is high with the majority having values higher than 0.98. Therefore, it can be said that almost all the districts of Rwanda experience similar rainfall patterns.
- For shorter distances, that is between 10 and 140, the correlation takes higher values than for distances from 150. From distance values of 150, the correlation values start to decrease with distance.

Scatter plot with fitted model



Insights

- The fitted model has a slope that is approximately flat negative slope. This indicates that the correlation between rainfall values of districts decreases only slightly as distance increases. This is evidence of related rainfall patterns in all Rwanda's 30 districts. Therefore, it can conclude that Rwanda has relatively uniform weather conditions in all its districts.

The estimated decay constant and C0

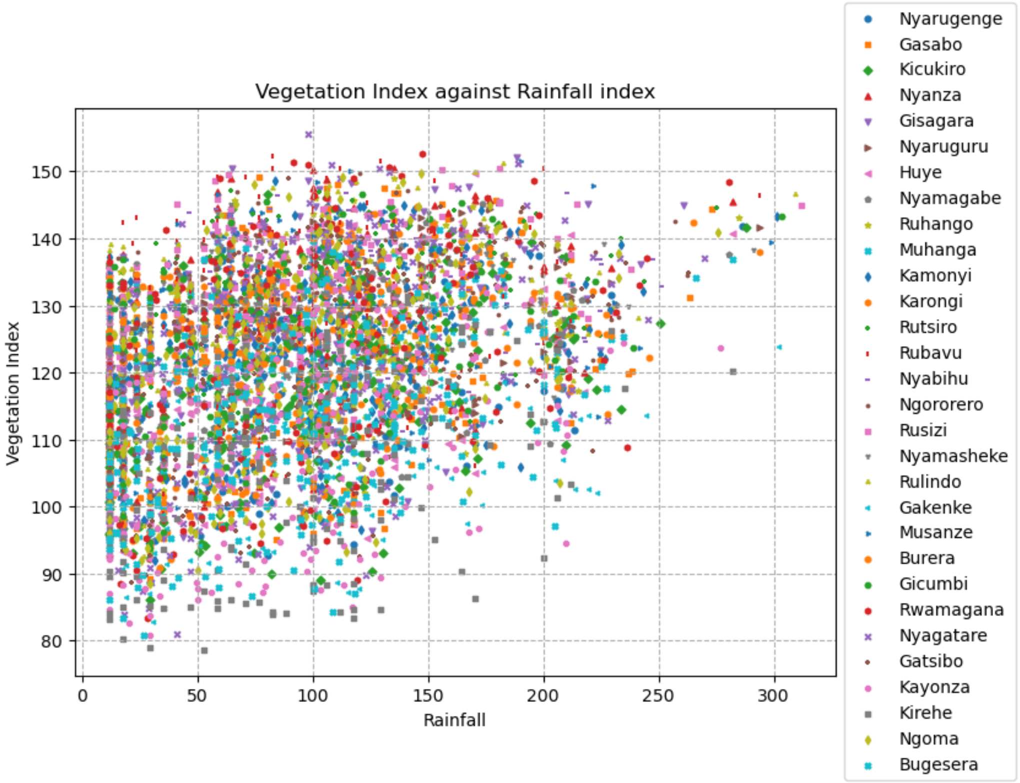
```
Estimated C0 and decay constant values are (0.993766113430371,
1.9072843363031717e-05) respectively
```

Comparative Analysis of Rainfall and Vegetation Index Time Series

Procedure

- To synchronize the rainfall and vegetation index data, the rainfall and vegetation index data frames were merged in an ‘inner’ manner using pandas’ merge function.
- Each of the 30 districts’ rainfall and vegetation index values were retrieved and plotted on a scatter plot using matplotlib’s scatter function.

Results



Insights

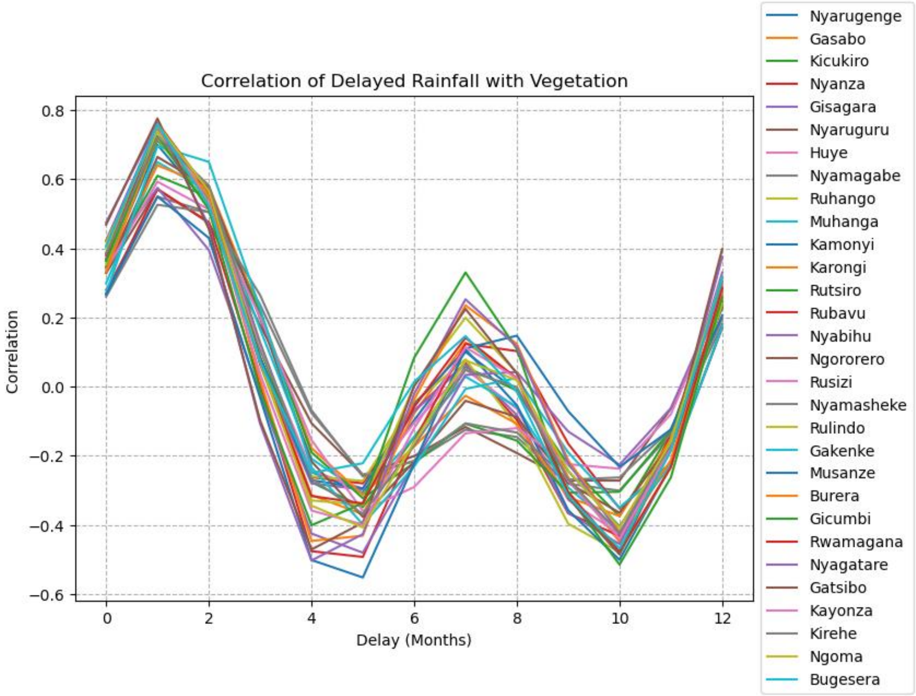
- The plot has most of the points concentrated between values 10 and 200 on the x-axis, and values 80 and 150 on the y-axis.
- Though not very clearly, it can be spotted that as rainfall increases, the vegetation index also increases.

Optimizing Time Delays for Predicting Vegetation Indices from Rainfall Data

Procedure

- With a range of delay values from 0 to 12 specified, the rainfall dataframe was shifted by each of the values in this range.
- The corresponding correlation between the delayed rainfall and vegetation index was calculated using the corr function.

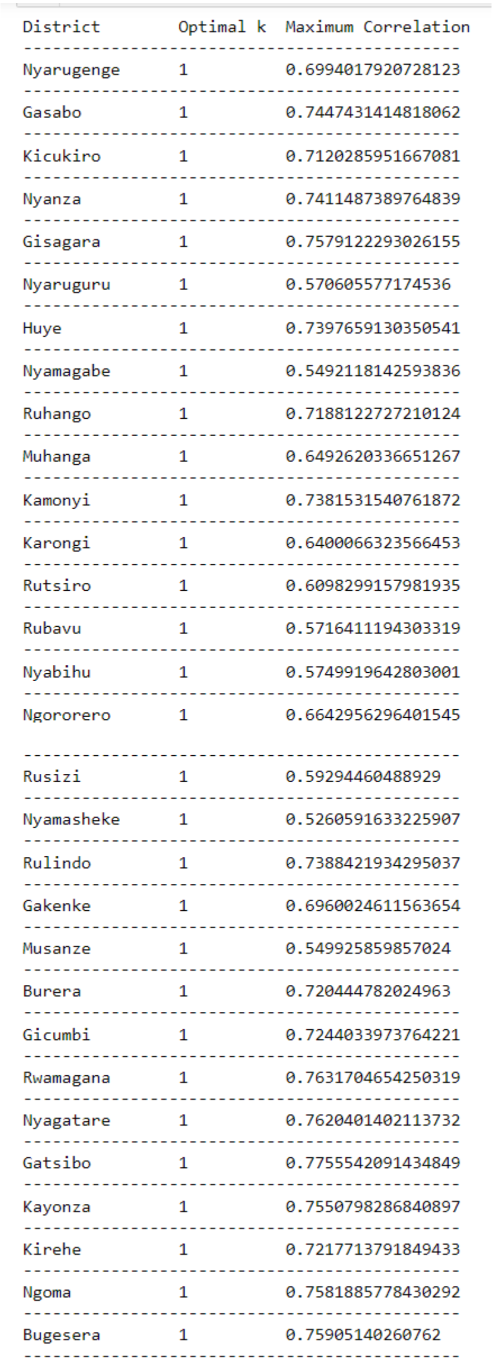
Results



Observation

- All districts have their highest correlation when the delay value k is equal to1.

Optimal delay



```
District        Optimal k  Maximum Correlation
---------------------------------------------
Nyarugenge      1          0.6994017920728123
---------------------------------------------
Gasabo          1          0.7447431414818062
---------------------------------------------
Kicukiro        1          0.7120285951667081
---------------------------------------------
Nyanza          1          0.7411487389764839
---------------------------------------------
Gisagara        1          0.7579122293026155
---------------------------------------------
Nyaruguru       1          0.570605577174536
---------------------------------------------
Huye            1          0.7397659130350541
---------------------------------------------
Nyamagabe       1          0.5492118142593836
---------------------------------------------
Ruhango         1          0.7188122727210124
---------------------------------------------
Muhanga         1          0.6492620336651267
---------------------------------------------
Kamonyi         1          0.7381531540761872
---------------------------------------------
Karongi         1          0.6400066323566453
---------------------------------------------
Rutsiro         1          0.6098299157981935
---------------------------------------------
Rubavu          1          0.5716411194303319
---------------------------------------------
Nyabihu         1          0.5749919642803001
---------------------------------------------
Ngororero       1          0.6642956296401545

---------------------------------------------
Rusizi          1          0.59294460488929
---------------------------------------------
Nyamasheke      1          0.5260591633225907
---------------------------------------------
Rulindo         1          0.7388421934295037
---------------------------------------------
Gakenke         1          0.6960024611563654
---------------------------------------------
Musanze         1          0.549925859857024
---------------------------------------------
Burera          1          0.720444782024963
---------------------------------------------
Gicumbi         1          0.7244033973764221
---------------------------------------------
Rwamagana       1          0.7631704654250319
---------------------------------------------
Nyagatare       1          0.7620401402113732
---------------------------------------------
Gatsibo         1          0.7755542091434849
---------------------------------------------
Kayonza         1          0.7550798286840897
---------------------------------------------
Kirehe          1          0.7217713791849433
---------------------------------------------
Ngoma           1          0.7581885778430292
---------------------------------------------
Bugesera        1          0.75905140260762
---------------------------------------------
```

Insights

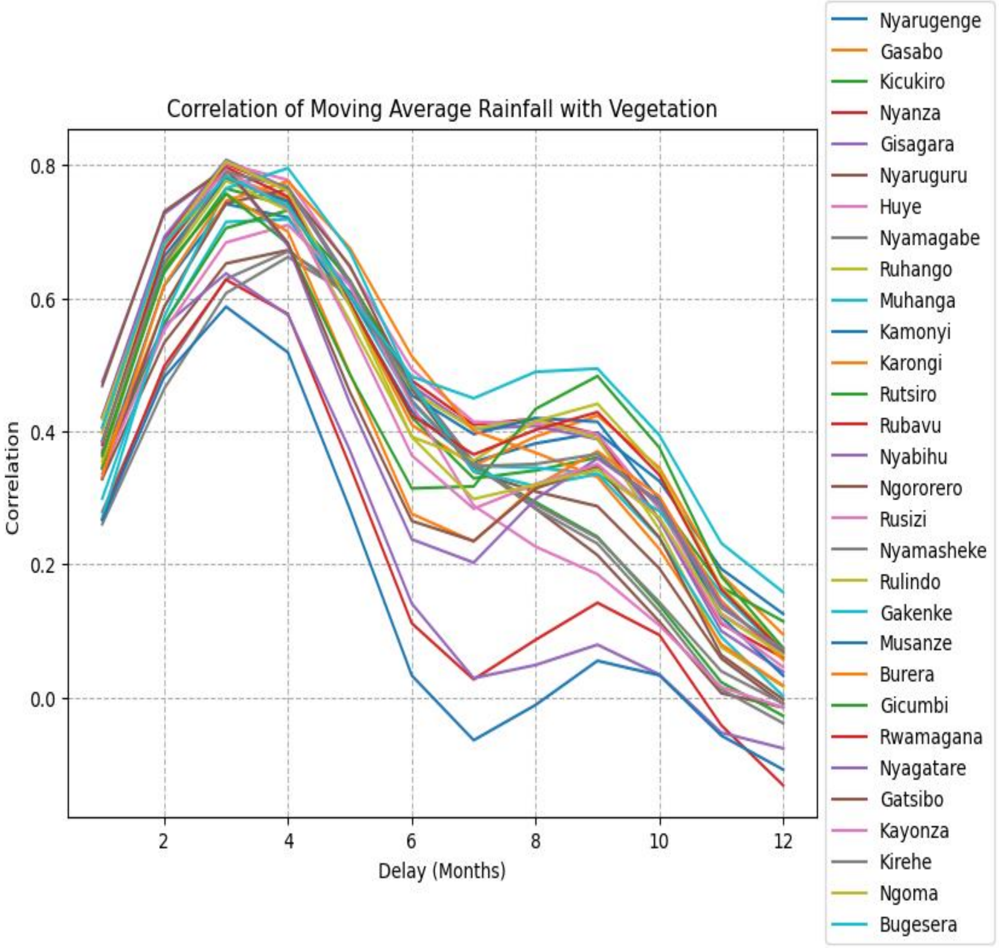
- All the districts have an optimal delay value of 1. This aligns with my intuition that the vegetation index of the current month is affected by the rainfall amount of the previous month.
- The correlation values are relatively high overall.
- Therefore, the optimal k is 1 for all districts and there is a consensus for the districts.

Moving Average Transformation of Rainfall for Enhanced Vegetation Index Prediction

Procedure

- Having defined window values from 1 to 12, the moving average rainfall was obtained by applying the rolling function on the rainfall data frame for each of these window values.
- The corresponding correlation between the SMA rainfall and vegetation index was calculated using the corr function during each iteration.

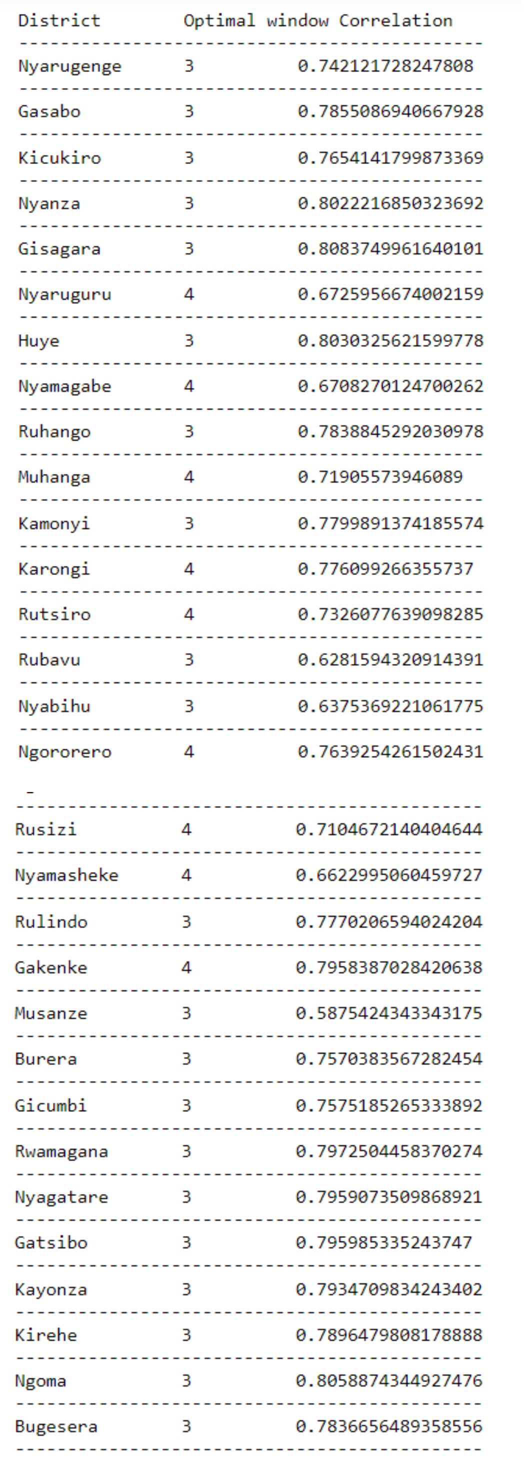
Results



Insights

- Most districts have an optimal window value of 3. Some districts have an optimal window value of 4.

Table



District	Optimal window	Correlation
Nyarugenge	3	0.742121728247808
Gasabo	3	0.7855086940667928
Kicukiro	3	0.7654141799873369
Nyanza	3	0.8022216850323692
Gisagara	3	0.8083749961640101
Nyaruguru	4	0.6725956674002159
Huye	3	0.8030325621599778
Nyamagabe	4	0.6708270124700262
Ruhango	3	0.7838845292030978
Muhanga	4	0.71905573946089
Kamonyi	3	0.7799891374185574
Karongi	4	0.776099266355737
Rutsiro	4	0.7326077639098285
Rubavu	3	0.6281594320914391
Nyabihu	3	0.6375369221061775
Ngororero	4	0.7639254261502431
Rusizi	4	0.7104672140404644
Nyamasheke	4	0.6622995060459727
Rulindo	3	0.7770206594024204
Gakenke	4	0.7958387028420638
Musanze	3	0.5875424343343175
Burera	3	0.7570383567282454
Gicumbi	3	0.7575185265333892
Rwamagana	3	0.7972504458370274
Nyagatare	3	0.7959073509868921
Gatsibo	3	0.795985335243747
Kayonza	3	0.7934709834243402
Kirehe	3	0.7896479808178888
Ngoma	3	0.8058874344927476
Bugesera	3	0.7836656489358556

Insights

- Most districts have an optimal window value of 3. Some districts have an optimal window value of 4.

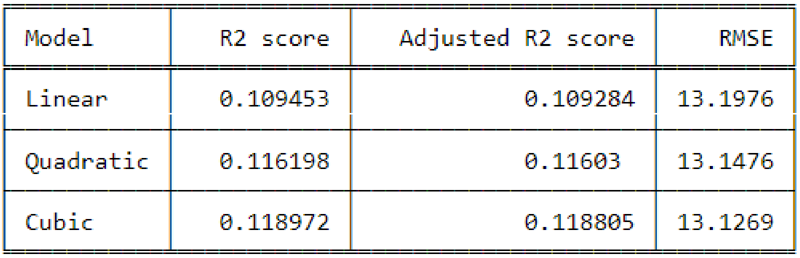
Modeling Vegetation Response to Rainfall Using Polynomial Regression

Procedure

- The metrics for each variable used were R2, adjusted R2 and root mean square error. These metrics were calculated with the help of sklearn's metrics module.
- For the delayed rainfall and moving average variables, their earlier obtained optimal values (i.e 1 and 3 respectively), were used.
- The linear, quadratic, and cubic models were fitted using NumPy's polyfit function.

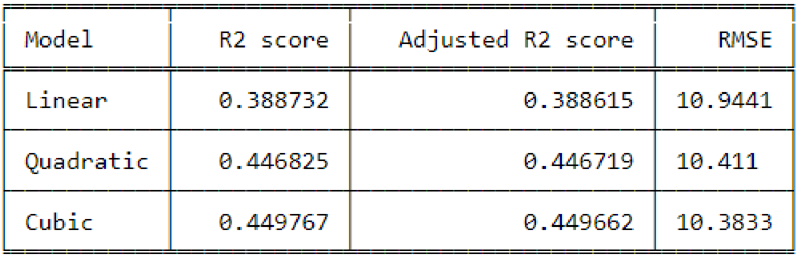
Results

Rainfall metrics values



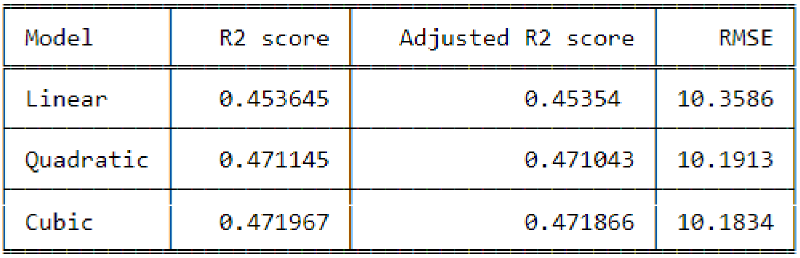
Model	R2 score	Adjusted R2 score	RMSE
Linear	0.109453	0.109284	13.1976
Quadratic	0.116198	0.11603	13.1476
Cubic	0.118972	0.118805	13.1269

Delayed Rainfall metrics values



Model	R2 score	Adjusted R2 score	RMSE
Linear	0.388732	0.388615	10.9441
Quadratic	0.446825	0.446719	10.411
Cubic	0.449767	0.449662	10.3833

SMA Rainfall metrics values



Model	R2 score	Adjusted R2 score	RMSE
Linear	0.453645	0.45354	10.3586
Quadratic	0.471145	0.471043	10.1913
Cubic	0.471967	0.471866	10.1834

Observations and Insights

- The rainfall variable has the lowest R-squared and adjusted R-squared values, and the highest RMSE generally, therefore this variable has the lowest performance among the three variables.
- For all variables, the cubic model is the best performing model.

- Overall, the moving average variable has the highest R-squared and adjusted R-squared values, and the lowest RMSE. Therefore, this variable is the best predictor variable for vegetation index among the three variables.

Is there any evidence for using a quadratic model to describe how the vegetation index varies with rainfall (or any of the above features: delayed rainfall and simple moving average rainfall)?

Answer

There is evidence of using a quadratic model to describe how vegetation index varies with rainfall for the delayed rainfall and the SMA rainfall variables. This is because of the high R-squared and adjusted R-squared obtained.

There is also evidence of using a cubic model the relationship between vegetation index and rainfall.

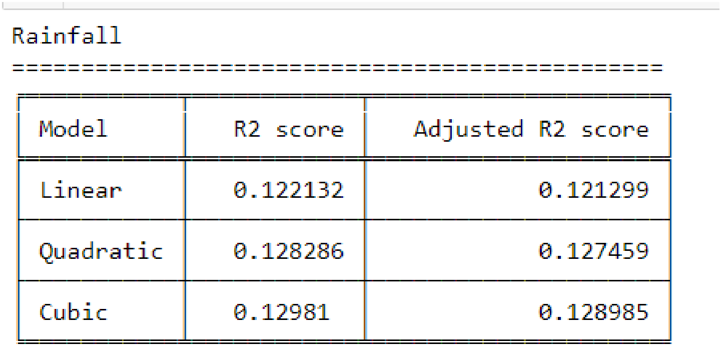
Cross-Validation of Rainfall Transformations for Vegetation Index Prediction

Procedure

- Using the optimal delay, and window for the delayed rainfall and SMA rainfall, the R-squared and adjusted R_squared values were obtained using cross validation by splitting the data into train and test sets, for linear, quadratic, and cubic models using these variables as predictor variables and vegetation index as the dependent variable.
- The above metrics were also obtained for the delayed SMA rainfall as the predictor variable for vegetation index. The delayed SMA was obtained by delaying the SMA by the optimal delay of 1 that was earlier obtained.

Results

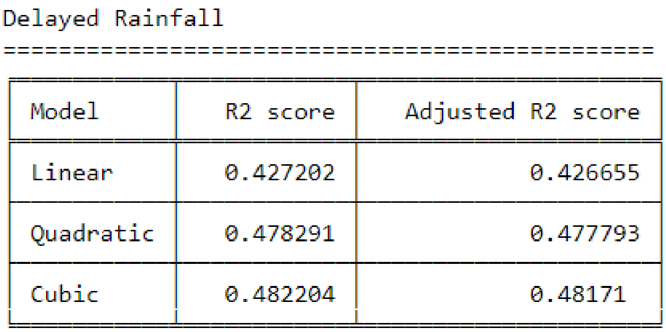
Rainfall



Rainfall
==

Model	R2 score	Adjusted R2 score
Linear	0.122132	0.121299
Quadratic	0.128286	0.127459
Cubic	0.12981	0.128985

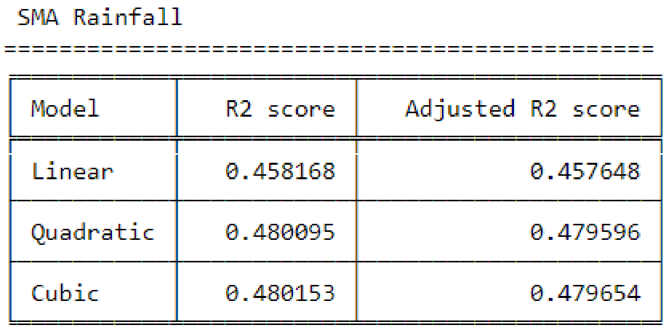
Delayed Rainfall



Delayed Rainfall
==

Model	R2 score	Adjusted R2 score
Linear	0.427202	0.426655
Quadratic	0.478291	0.477793
Cubic	0.482204	0.48171

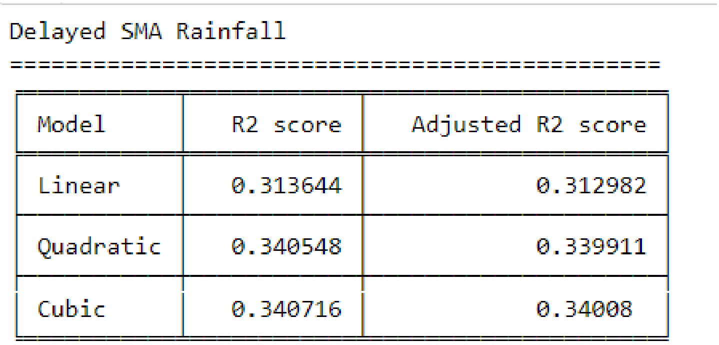
SMA Rainfall



```
 SMA Rainfall
===============================================
```

Model	R2 score	Adjusted R2 score
Linear	0.458168	0.457648
Quadratic	0.480095	0.479596
Cubic	0.480153	0.479654

Delayed SMA



```
Delayed SMA Rainfall
===============================================
```

Model	R2 score	Adjusted R2 score
Linear	0.313644	0.312982
Quadratic	0.340548	0.339911
Cubic	0.340716	0.34008

Insights

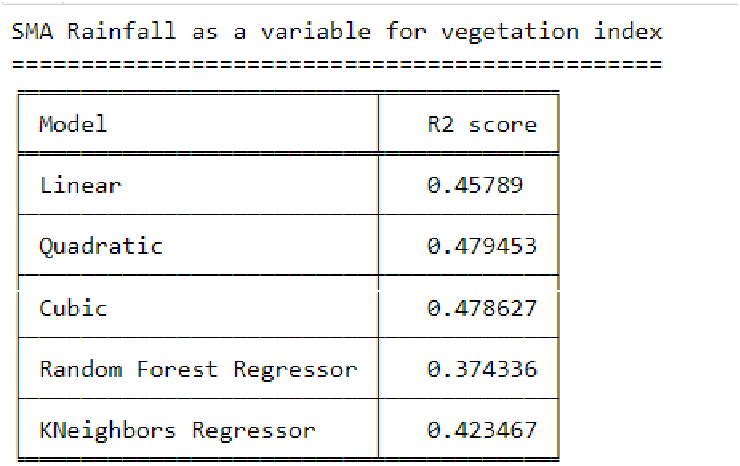
- The SMA rainfall is the best predictor variable for the vegetation index among all the 4 variables since it has the highest R-squared and adjusted R-squared values for both the linear and quadratic models.
- In all the variables, the cubic model is the best performing fitted model.

Model Selection and Evaluation for Predicting Vegetation Index Using Rainfall Data

Procedure

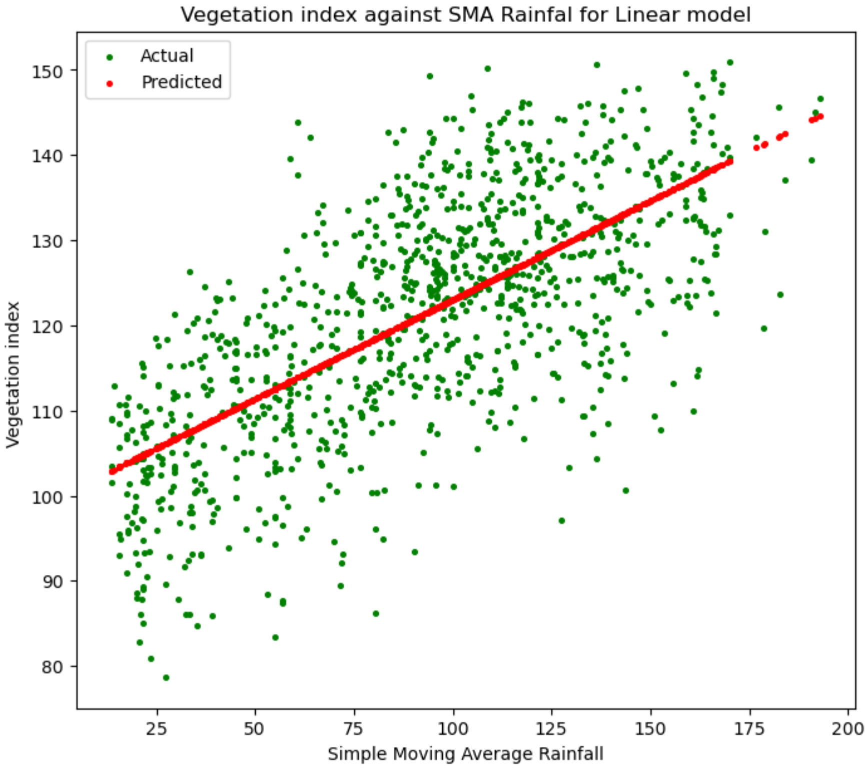
- The above best obtained variable for predicting the vegetation index was the SMA rainfall variable. In addition to the linear, quadratic, and cubic models, random forest regressor and KNeighbors regressor.
- All the models were cross validated using sklearn's train_test_split function, and sklearn's r2_score function for calculating their corresponding R-squared values.

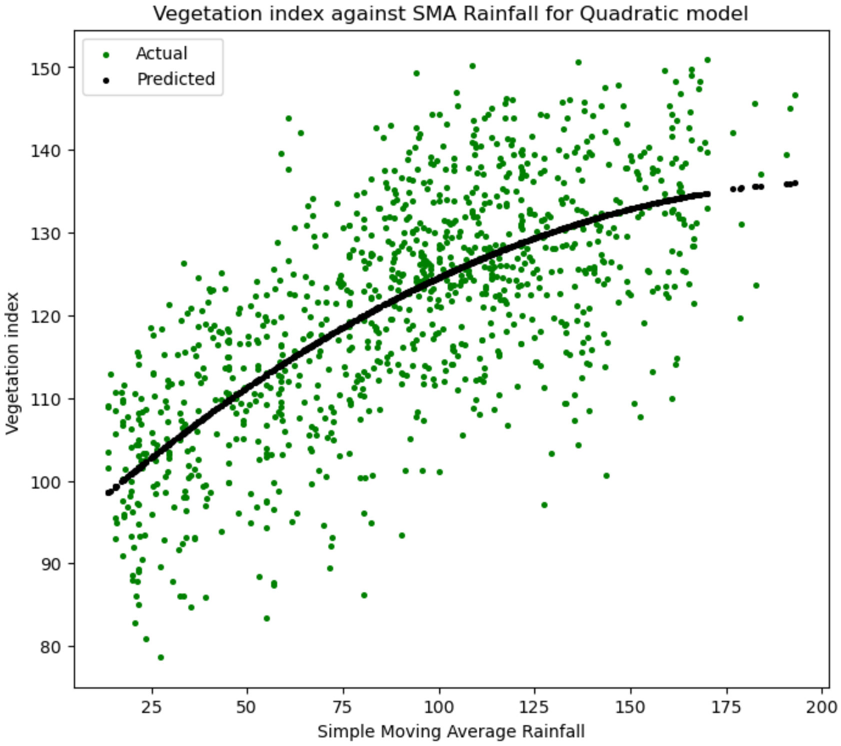
Results

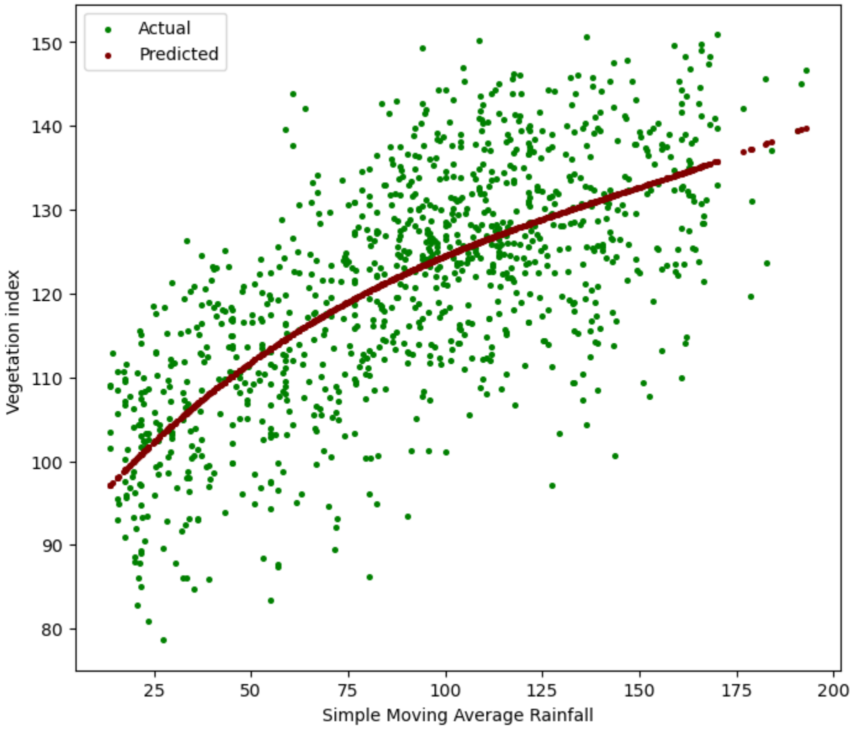


```
SMA Rainfall as a variable for vegetation index
===============================================
```

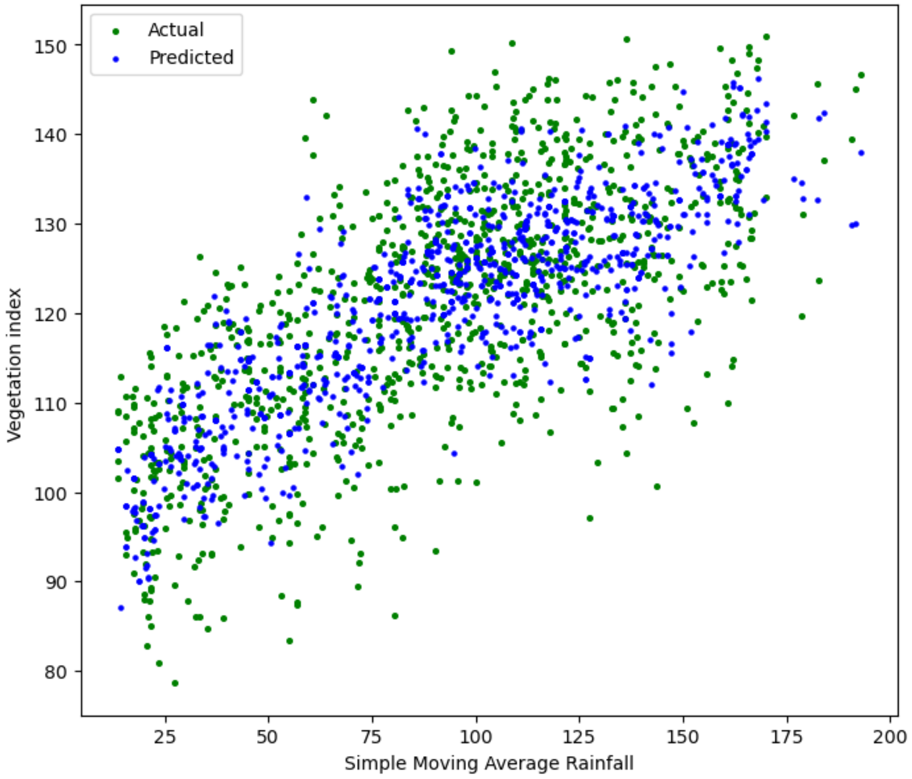
Model	R2 score
Linear	0.45789
Quadratic	0.479453
Cubic	0.478627
Random Forest Regressor	0.374336
KNeighbors Regressor	0.423467


Vegetation index against SMA Rainfal for Linear model
Actual
Predicted
Vegetation index
150
140
130
120
110
100
90
80
25
50
75
100
125
150
175
200
Simple Moving Average Rainfall

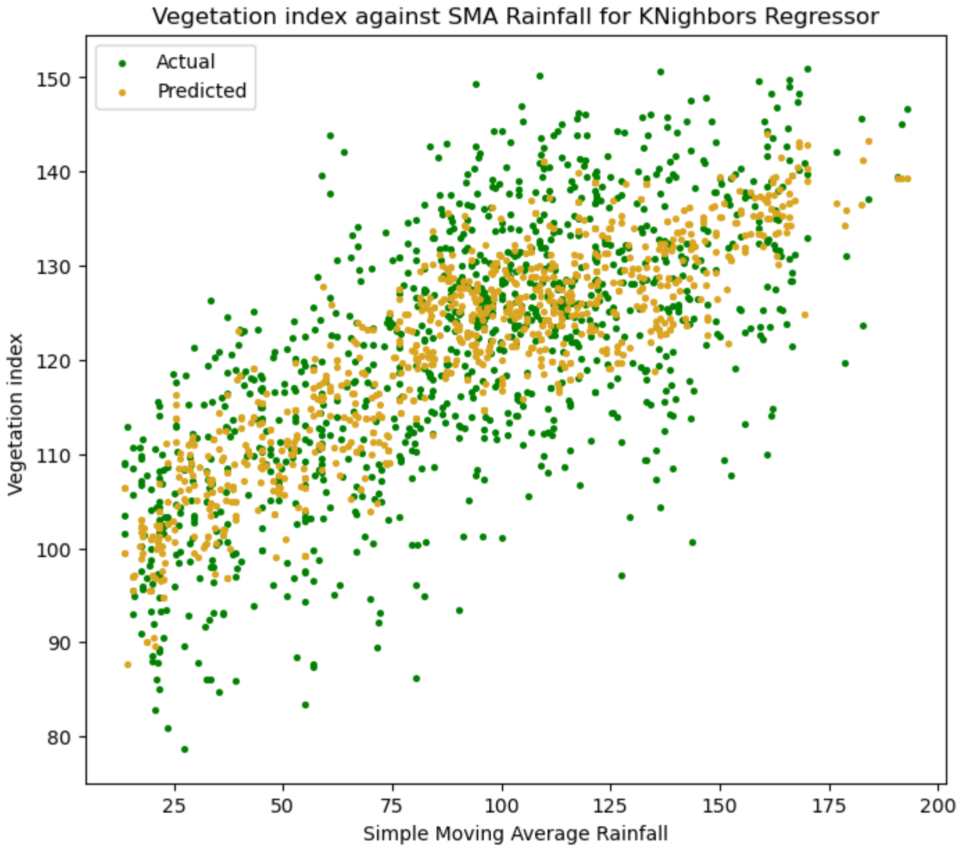

Vegetation index against SMA Rainfall for Quadratic model
Actual
Predicted
Vegetation index
150
140
130
120
110
100
90
80
25
50
75
100
125
150
175
200
Simple Moving Average Rainfall



Vegetation index against SMA Rainfall for cubic model



Vegetation index against SMA Rainfall fro Random Forest Regressor



Insights

- The quadratic model has the highest R-squared value among all the 5 models. The cubic model is the second-best model. The least performing model is the random forest regressor.

Optimal model that I would recommend for predicting the vegetation index

Answer

From the above model metrics, the quadratic model is the best performing model with the highest R-squared value among all other models. Therefore, I recommend the quadratic model for predicting vegetation index with SMA rainfall as the predictor variable.

REFERENCES

[1] “Rwanda Weather & Climate (+ Climate Chart),” SafariBookings.com. Accessed: Mar. 25, 2024. [Online]. Available: https://www.safaribookings.com/rwanda/climate