Ocean Protocol ::
Dimitra Bounty Phase 2

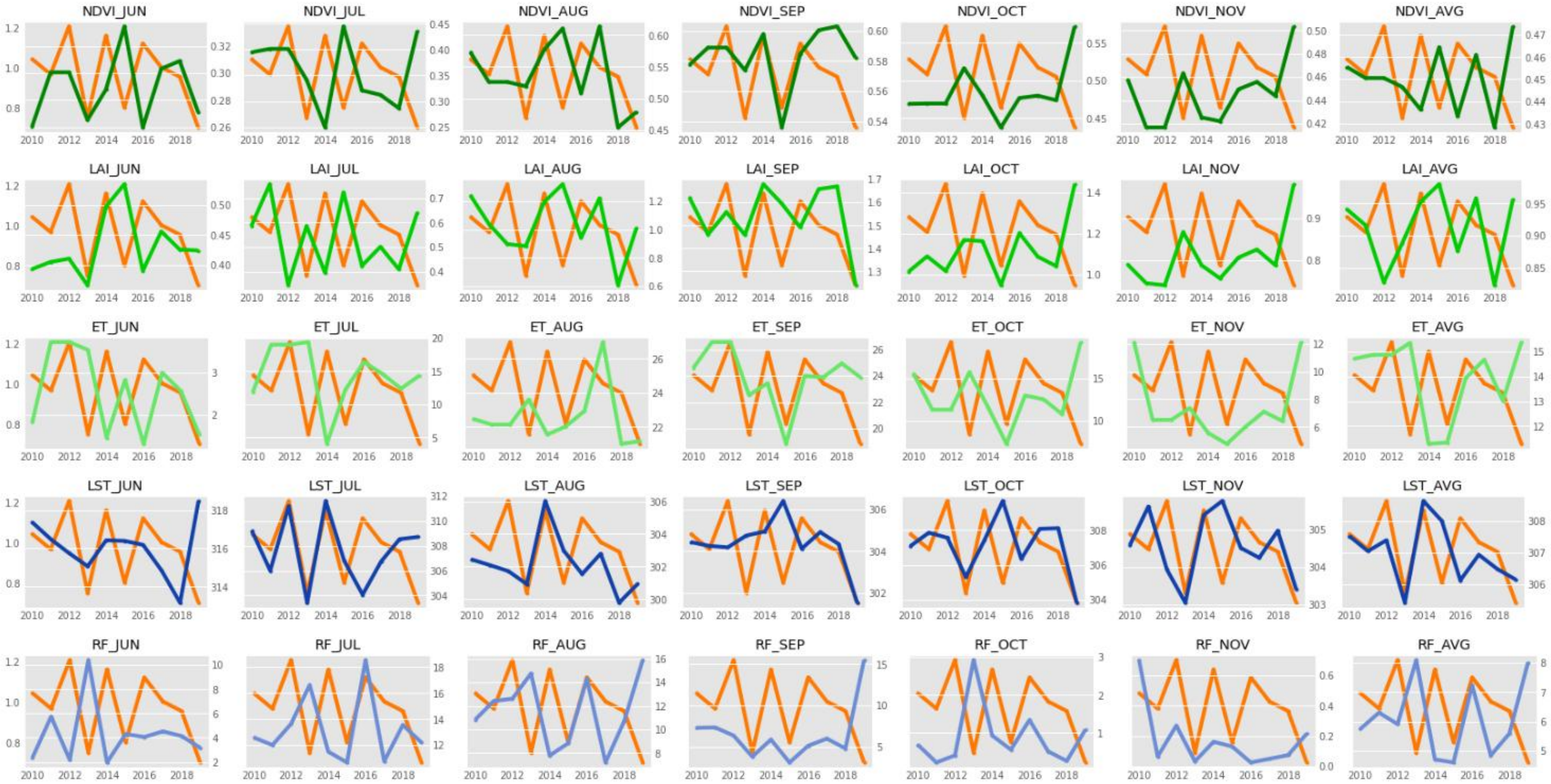
DIMITRA ocean

Yearly features values

On the picture below we shown the comparison of target values (Yearly yield of soybean) with every feature in the dataset:

- Yearly yield of soybean is depicted as orange line on each plot and their values are on the *LEFT Y* axis.
- Indexes is depicted as green lines (NDVI, LAI, ET), Land Surface Temperature and Rainfall (LST, RF) as blue lines and their values are on the *RIGHT Y* axis.
- Year is on the X axis.

Then we can visually compare the dependence of each feature with the target by analyzing the differences of the values from year to year.



Comparing yearly differences features and target values

We compared the yearly feature differences with the target differences and set new variable with the following values:

- 1 – if the feature and target are going the same way (increasing or decreasing at the same time);
- 0 – if the feature or target is not changes significantly;
- -1 – if the feature and target are going the opposite way.

Then we calculated the sum of that variable and created the *Score* variable that is showing the similarity between feature and target. The results are shown in the right table. We found the following interesting facts from that table:

- Yearly average values of all indexes (NDVI, LAI, ET) and Rainforest have negative correlation with the target.
- Land Surface Temperatures have a positive correlation with the target.
- All indexes have a positive correlation with the target in September.

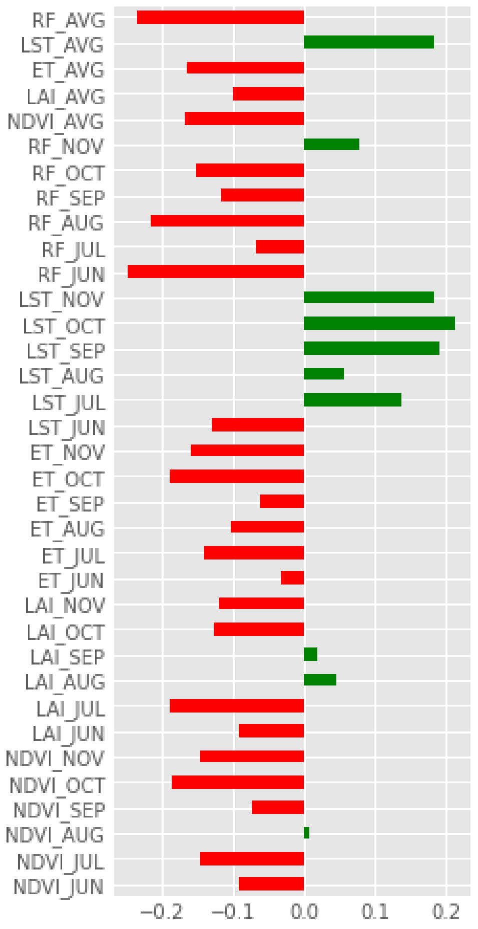
Based on that table and univariate feature selection methods we will select the most useful features for a simple model for predicting the target value.

Feature	Month	Difference									Score
NDVI	JUN	-1	0	1	1	-1	-1	-1	-1	1	-2
	JUL	0	0	1	-1	-1	-1	1	1	-1	-1
	AUG	1	0	0	1	-1	-1	-1	1	-1	-1
	SEP	-1	0	1	1	1	1	-1	-1	1	2
	OCT	0	0	-1	-1	1	1	0	0	-1	-1
	NOV	1	0	-1	-1	0	1	0	1	-1	0
	AVG	1	0	0	-1	-1	-1	-1	1	-1	-3
LAI	JUN	-1	1	1	1	-1	-1	-1	1	0	0
	JUL	-1	-1	-1	-1	-1	-1	-1	1	-1	-7
	AUG	1	-1	0	1	-1	-1	-1	-1	-1	-4
	SEP	1	1	1	1	1	-1	-1	0	1	4
	OCT	-1	-1	-1	0	1	1	1	1	-1	0
	NOV	1	0	-1	-1	1	1	-1	1	-1	0
	AVG	1	-1	-1	1	-1	-1	-1	1	-1	-3
ET	JUN	-1	0	0	-1	-1	-1	-1	1	1	-3
	JUL	-1	0	0	-1	-1	1	1	1	-1	-1
	AUG	1	0	-1	-1	-1	1	-1	-1	0	-3
	SEP	-1	0	1	1	1	1	0	-1	1	3
	OCT	1	0	-1	-1	1	1	0	1	-1	1
	NOV	1	0	-1	-1	1	1	-1	1	-1	0
	AVG	0	0	-1	-1	0	1	-1	1	-1	-2
LST	JUN	1	-1	1	1	0	0	1	1	-1	3
	JUL	1	1	1	1	1	-1	-1	-1	0	2
	AUG	1	-1	1	1	1	-1	-1	1	-1	1
	SEP	0	0	-1	0	-1	-1	-1	1	1	-2
	OCT	-1	-1	1	1	-1	-1	-1	0	1	-2
	NOV	-1	-1	1	1	-1	-1	1	-1	1	-1
	AVG	1	1	1	1	1	-1	-1	1	1	5
RF	JUN	-1	-1	-1	-1	-1	0	0	0	1	-4
	JUL	1	1	-1	-1	1	1	1	-1	1	3
	AUG	-1	0	-1	-1	-1	1	1	0	-1	-3
	SEP	0	0	1	1	1	1	-1	-1	-1	1
	OCT	1	1	-1	-1	1	1	1	1	-1	3
	NOV	1	1	1	1	1	-1	0	0	-1	3
	AVG	-1	-1	-1	-1	1	1	1	-1	-1	-3

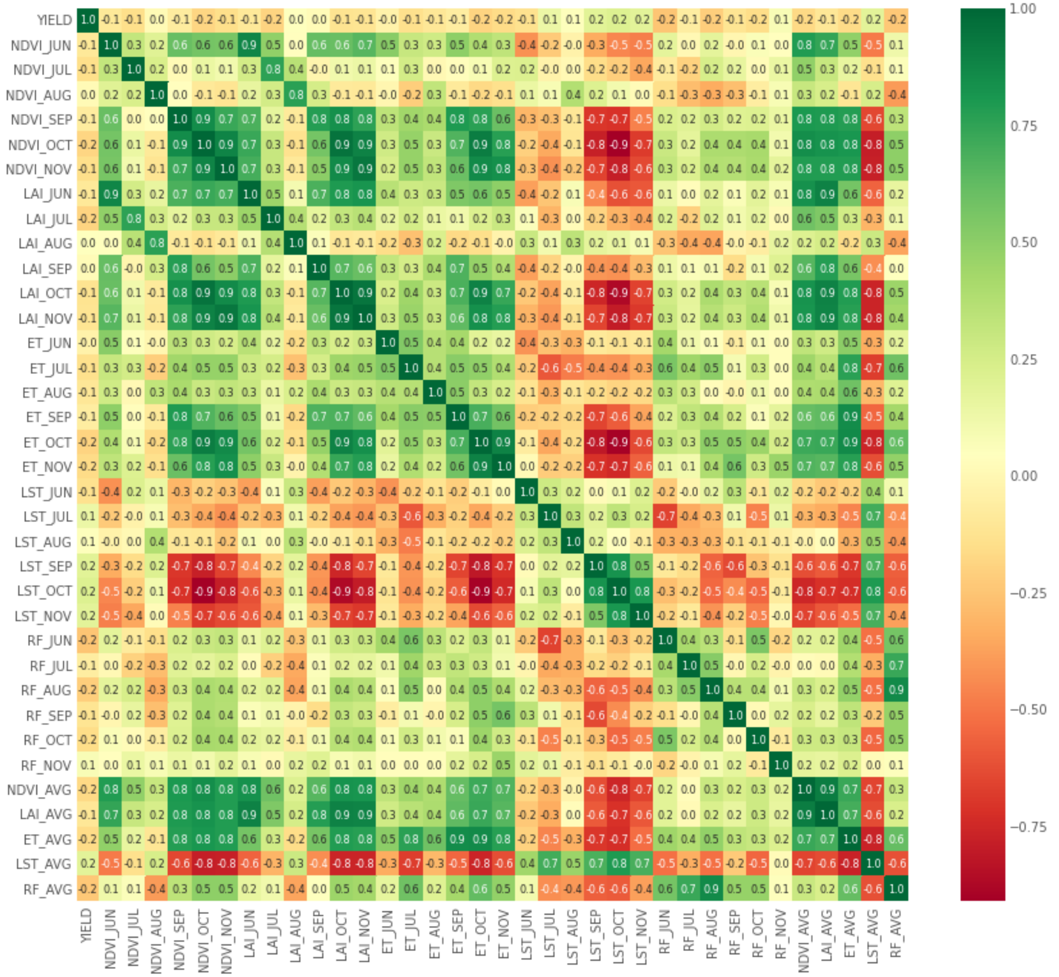
DIMITRA ocean

Feature correlations

Correlations with target:



Correlation matrix:



- We received a similar results comparing with previous table.
- Also we have checked statistical tests (ANOVA, Mann-Whitney, T-test) for comparing mean values between two or more groups and can remove some features.
- For our simple model we selected the most useful features: LST_OCT, RF_JUN (these features are not correlated with each other) and will build simple linear models to predict target value: Yi= a*Xi + b. Final prediction is the average of them.

Modeling phase description

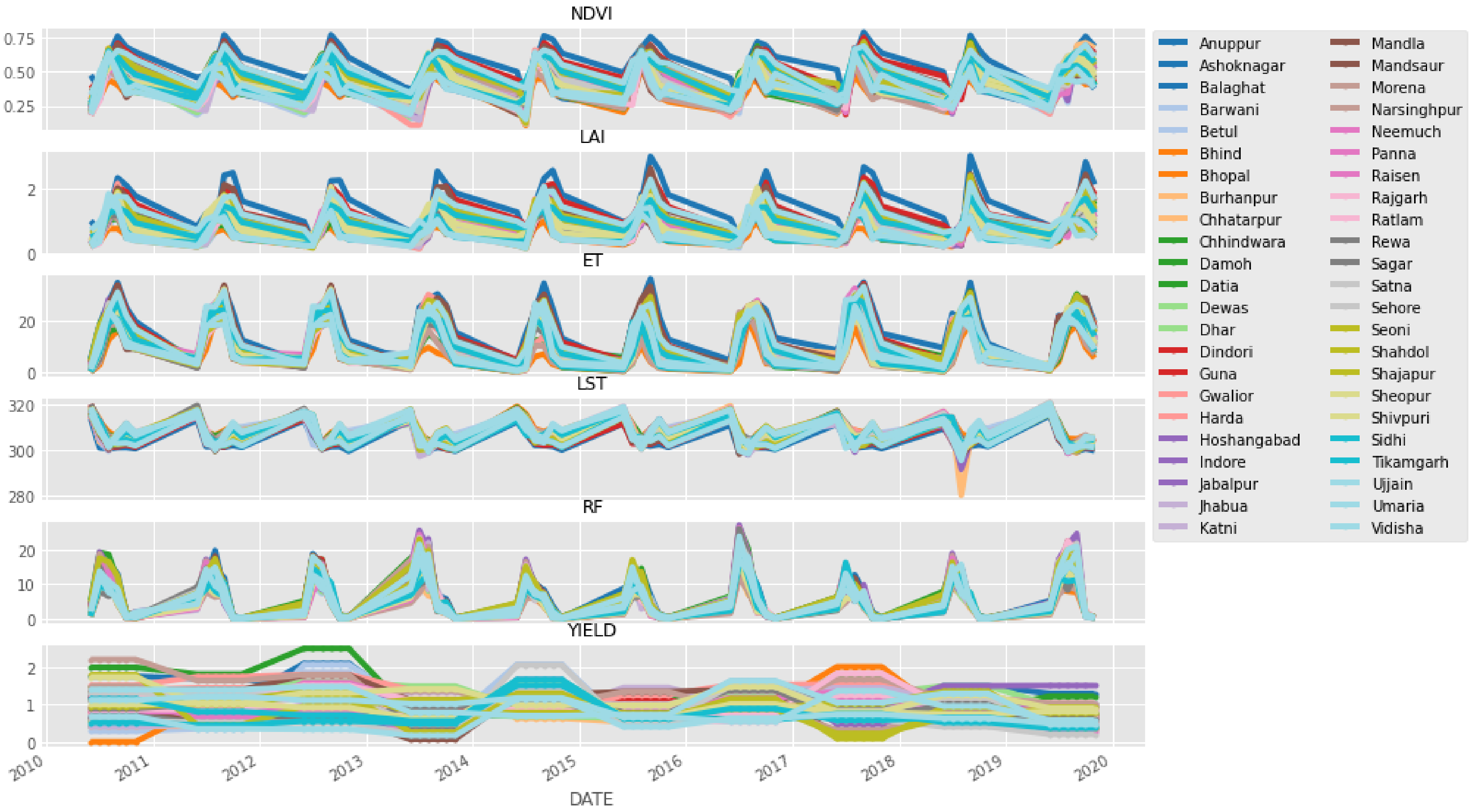
1. For models evaluation score we chose *mean absolute error*.
2. Validation set consists of the data of randomly selected 10 districts.
3. We have repeated experiments 5 times for different validation samples and calculated average and standard deviation of the scores.
4. For the linear models we preprocessed the data by applying Standard Scaler. So all the features have been transformed to z-scores.
5. Final scores shown in the table below.

Experiment number	Simple model	Linear model	Xgboost	Random Forest
1	0,3434	0,3194	0,3192	0,3213
2	0,3139	0,3295	0,3183	0,3169
3	0,2985	0,2633	0,2649	0,2526
4	0,3324	0,3176	0,3046	0,3087
5	0,3183	0,2967	0,3348	0,2994
AVERAGE	0,3213	0,3053	0,3083	0,2997
STD	0,01728	0,02633	0,02654	0,02767

*More details you can find in the following script: *1_Feature_Selection.ipynb*

DIMITRA ocean

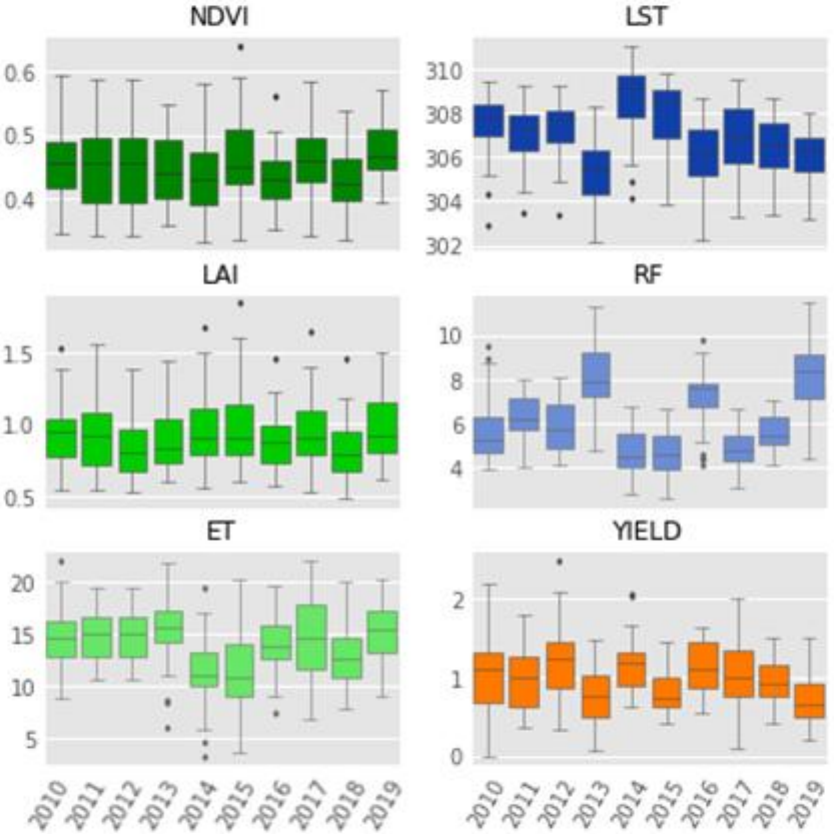
Time series districts plots for every feature



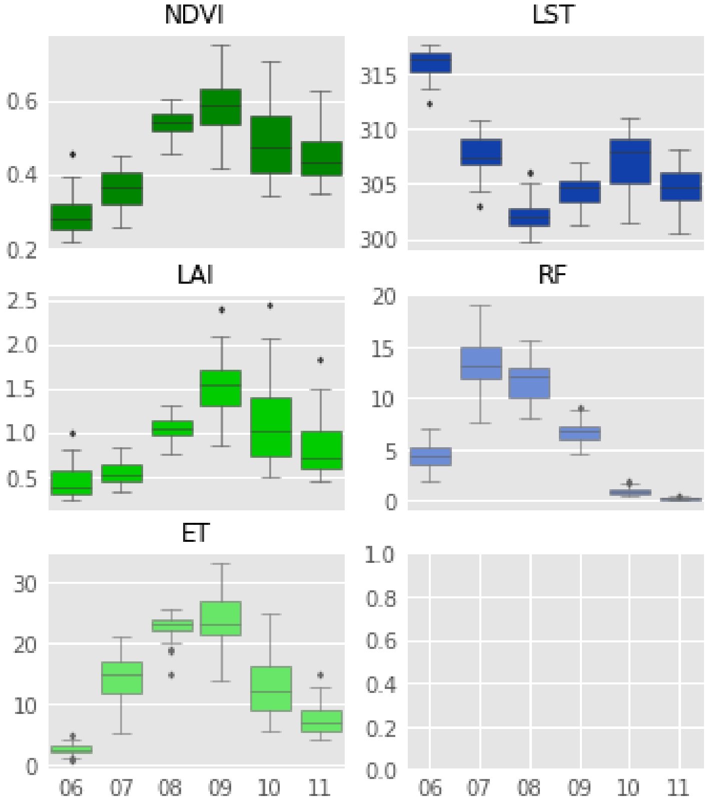
On the picture above we can see all points in the dataset and it is hard to see useful facts from that. So on the next slide we will build distributions by years and months and show it on the box-plots. I will be easier to get some ideas from that.

Yearly and monthly box-plots for every feature

Yearly box-plots:



Monthly box-plots:



- For the target feature Soybean yield there is no positive trend unlike wheat yield (it was analyzed on the previous phase).
- It seems that Rainforest have repeating similar pattern every 3 years: high – low – low.
-If Rainforest is growing then Land Surface Temperatures is falling.
- All indexes are highly correlated with each other.

- For all indexes, we see a similar picture. The highest value is reached in September. In October, the largest variance due to the fact that the index falls after the peak.
- The highest value of Rainforest is in July.
- Land Surface Temperatures have the highest value in June.