Ocean Protocol :: Dimitra Bounty Phase 2

Challenge Dataset

The dataset provided for this challenge can be downloaded at no cost on the Ocean Market (Polygon Network) by accessing the following link: https://bit.ly/dimitra-p2.

It provides crop yield data for soybean in the 46 districts of the Madhya Pradesh State, India for a period of 10 years, between 2010-2019.

It also provides MODIS satellite data for the same 10 year period, for the months June - November, for the following features:

NDVI: Normalized Difference Vegetation Index
LAI: Leaf Area Index
ET: Evapotranspiration
LST: Land Surface Temperature
RF: Rainfall

Note: The soybean crop duration in Madhya Pradesh is from June to November hence the data extracted for the variables is from June to November month.

Challenge

Given the (target) soybean yields and the satellite data (features), we ask you to:
- Rank the factors that affect yields and build a simple model to predict yields based on the features.
- Identify and explain any trends in the yields over the 10 year period.

Recommended Procedures
You MUST AT LEAST perform the graphical methods **described below** in order to be considered for the competition.

Understand the data
Learn the meaning of the satellite data features.
Understand that the satellite data features not only include information about soybean crops but also about any other vegetation, soil and/or land-cover/land-use present in the districts.

Feature Selection Analysis
Graphical Method: Based on graphs of the features (NDVI, LAI, ET, LST, RF) values versus time and of the target (yield) values versus time, can you rank (from most positive to most negative) the correlation between each feature and the yield? Can you rank the features correlations between each other?
Potential Advanced Methods:** Filter Methods (Univariate selection (ANOVA), Chi Square, based on correlation, based on variance); Wrapper Methods (Forward Selection, Backward Selection, Exhaustive Search); Embedded Methods (Lasso/Ridge (using ElasticNet), Tree based selection, Regression coefficients), Hybrid Methods (Feature shuffling, Recursive feature elimination, Recursive feature addition), PCA.

Time Series Analysis
Graphical Method: In the above graphs, can you identify any seasonal (intra-year), cyclical and/or secular (long-term) trends? Do the intra-year shapes of the curves change from year to year?(compare the slopes and curvatures of curves for the various years). What are the reasons and consequences of the variations and trends?
Potential Advanced Methods:** To develop a model to predict future yields consider ARIMA, Prophet, Vector Autoregression, Catboost Regressor, any regressor.
** Feel free to augment the challenge dataset (see below) with any open source real-data dataset of your choice. If you do so, please reference the data and provide a link to it. **Bonus points will be awarded for publishing the referenced data on the Ocean Market!!!**

Evaluation

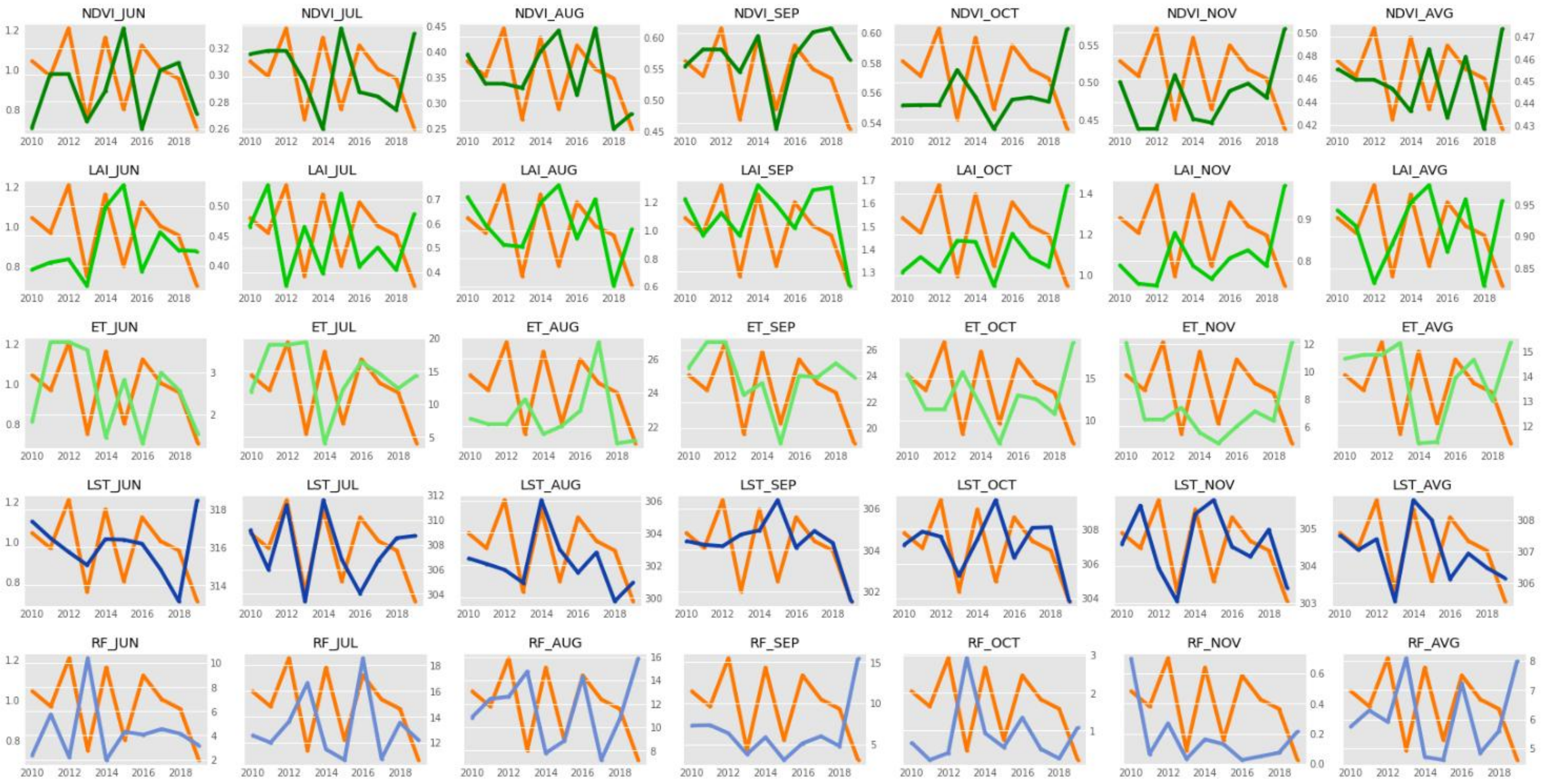
A panel of evaluators from Ocean and Dimitra will independently review and rank submission entries selecting 1st, 2nd and 3rd place winners. The selection will be announced publicly and the community will be invited to vote for the *Community Choice* Award winner.

Evaluation Criteria & Scoring
Minimum requirement: Plotting all the features (NDVI, LAI, ET, LST, RF) and the yearly yields versus time (10 years range).

Feature Selection Analysis:
15% - Performing the Graphical Method for the Feature Selection Analysis and interpreting the results (textual - no code required).
10% - Augmenting the data, specifically, finding and using additional open source datasets of relevant real data to complement the given data (performing any valuable data preprocessing, if necessary).
10% - Deriving a simple empirical formula to predict yields as a function of the features, if possible from your above Graphical Method analyses or otherwise.
15% - Training a machine learning model to calculate/predict yields based on features data.
Providing a file with your code with comments explaining your pipeline's flow.
Explaining your criteria for dealing with any of the following items, whenever applicable: selected algorithm(s), hyperparameter values, ensemble method, cross validation, method for determining your model's accuracy, and others.

Time Series Analysis:
15% - Performing the Graphical Method for the Time Series Analysis and interpreting the results (textual - no code required).
10% - Augmenting the data, specifically, finding and using additional open source datasets of relevant real data to complement the given data (performing any valuable data preprocessing, if necessary).
10% - Deriving a simple empirical formula to predict future seasons yields as a function of features and yields data, if possible from your above Graphical Method analyses or otherwise.
15% - Training a machine learning model to predict future seasons yields based on features and yields data.
Providing a file with your code with comments explaining your pipeline's flow.
For unfinished work partial credit will apply.

Yearly features values



On the picture above 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:

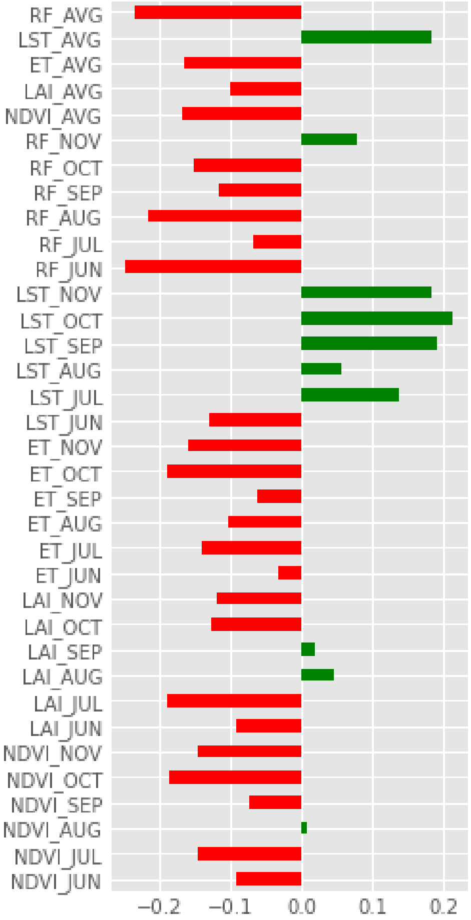
- Yearly average values of all indexes (NDVI, LAI, ET) and Rainforest have negative correlation with the target.
- Land Surface Temperatures have a positive correlation with the target.
- All indexes have a positive correlation with the target in September.

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.

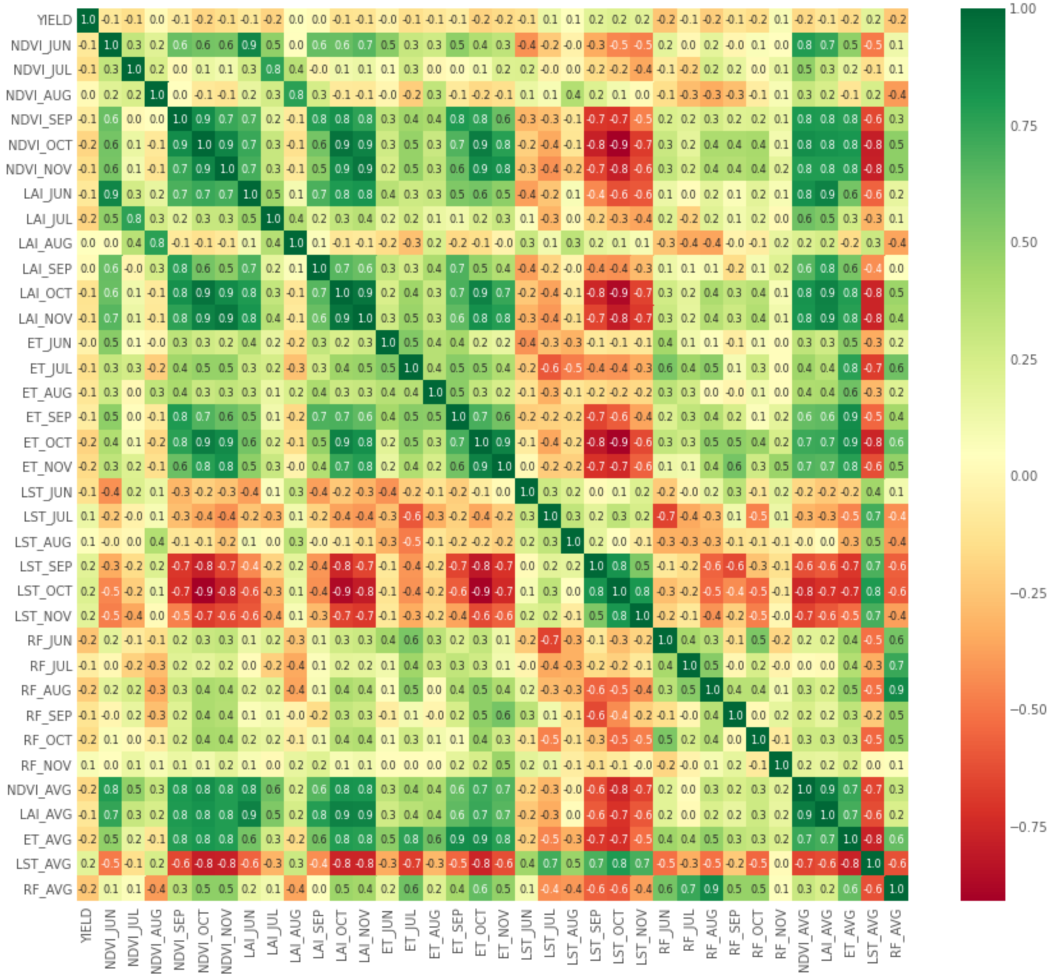
Feature	Month	Difference									Score
NDVI	JUN	-1	0	1	1	-1	-1	-1	-1	1	-2
	JUL	0	0	1	-1	-1	-1	1	1	-1	-1
	AUG	1	0	0	1	-1	-1	-1	1	-1	-1
	SEP	-1	0	1	1	1	1	-1	-1	1	2
	OCT	0	0	-1	-1	1	1	0	0	-1	-1
	NOV	1	0	-1	-1	0	1	0	1	-1	0
	AVG	1	0	0	-1	-1	-1	-1	1	-1	-3
LAI	JUN	-1	1	1	1	-1	-1	-1	1	0	0
	JUL	-1	-1	-1	-1	-1	-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

Feature correlations

Correlations with target:

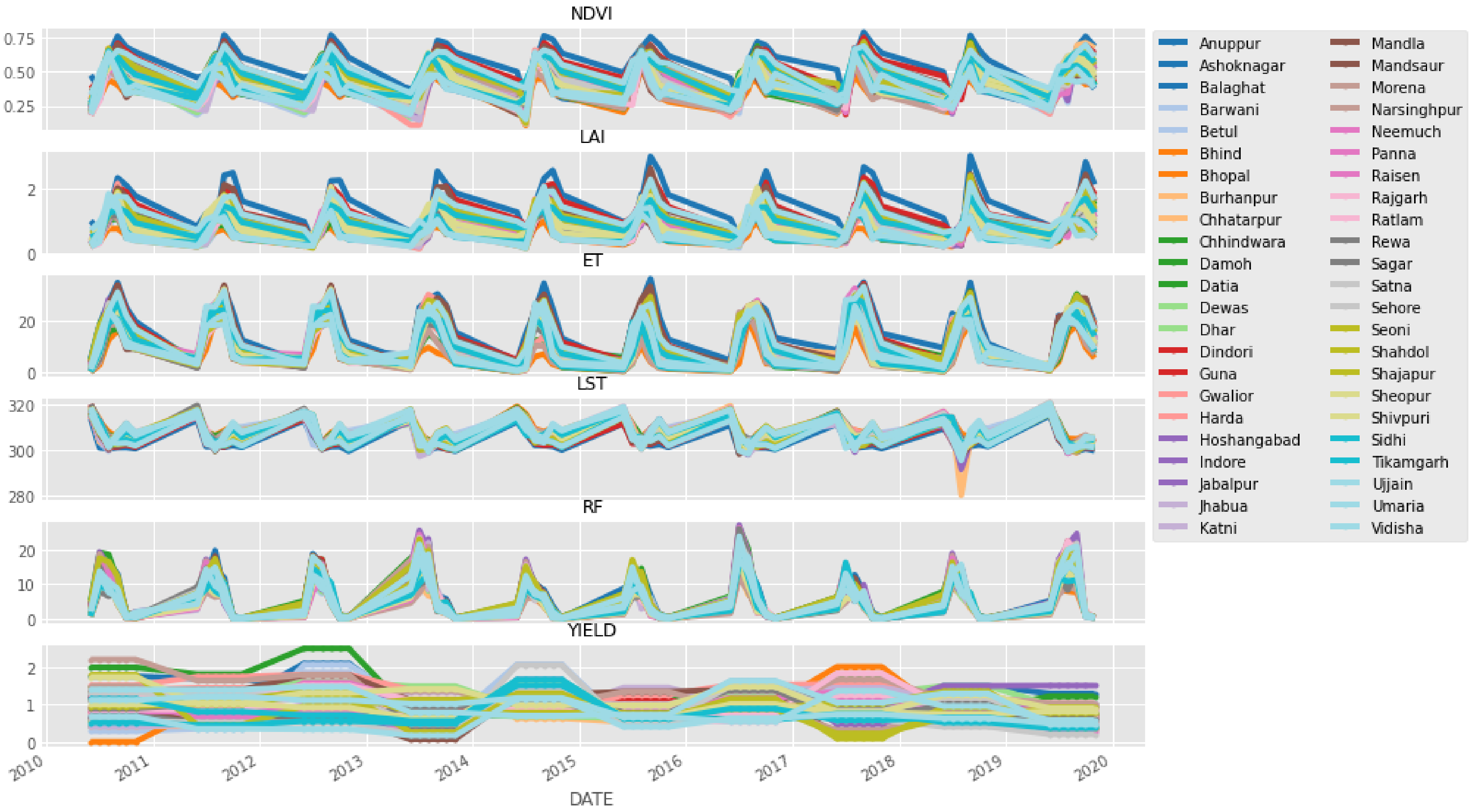


Correlation matrix:



- 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 reject null hypothesis for the following features: NDVI_AUG, NDVI_SEP, LAI_AUG, LAI_SEP, ET_JUN, ET_SEP, LST_AUG, RF_JUL, RF_NOV (we can drop these features).
- There are a lot of features that correlates with each so that to prevent multicollinearity problem we will remove them and select the best one.

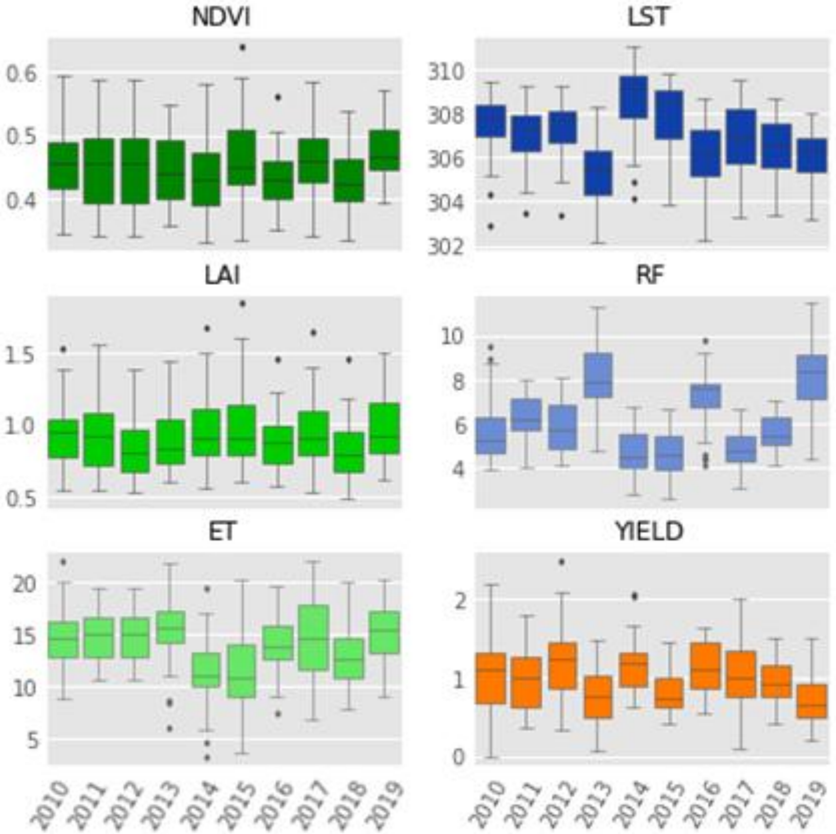
Time series districts plots for every feature



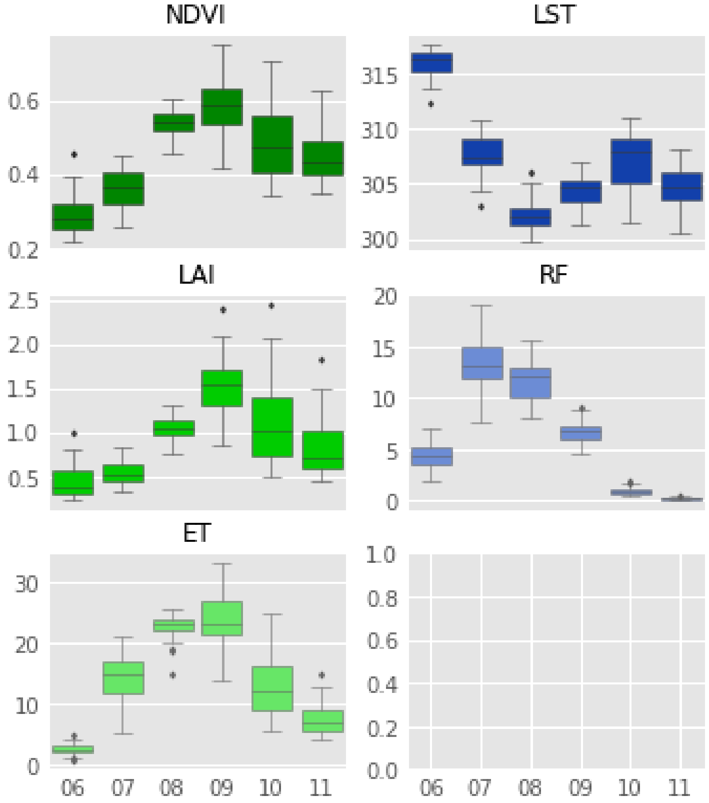
On the picture above we can see all points in the dataset and it is hard to see useful facts from that. So on the next slide we will build distributions by years and months and show it on the box-plots. I will be easier to get some ideas from that.

Yearly and monthly box-plots for every feature

Yearly box-plots:



Monthly box-plots:



TODO

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.

DIMITRA ocean