Ocean Protocol ::
Dimitra Bounty Phase 2

DIMITRA ocean

Yearly features values

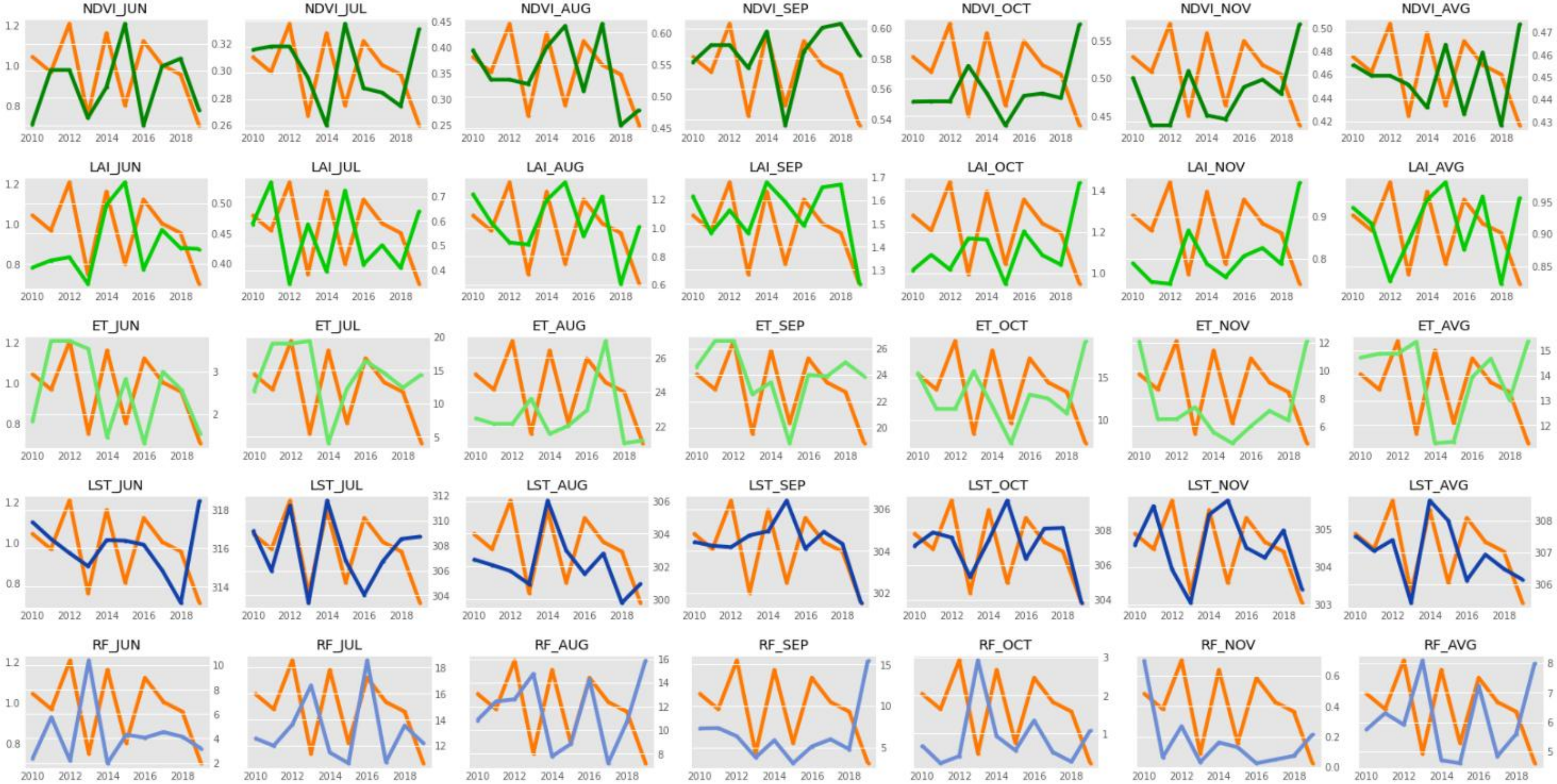
On the picture below we shown the comparison of target values (Yearly yield of soybean) with every feature in the dataset.

□ 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.

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.

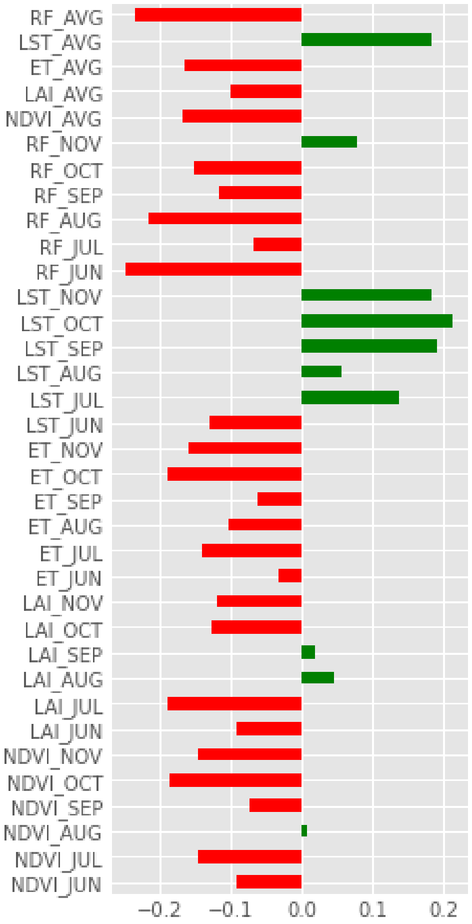
□ All indexes have a positive correlation with the target in September.

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.

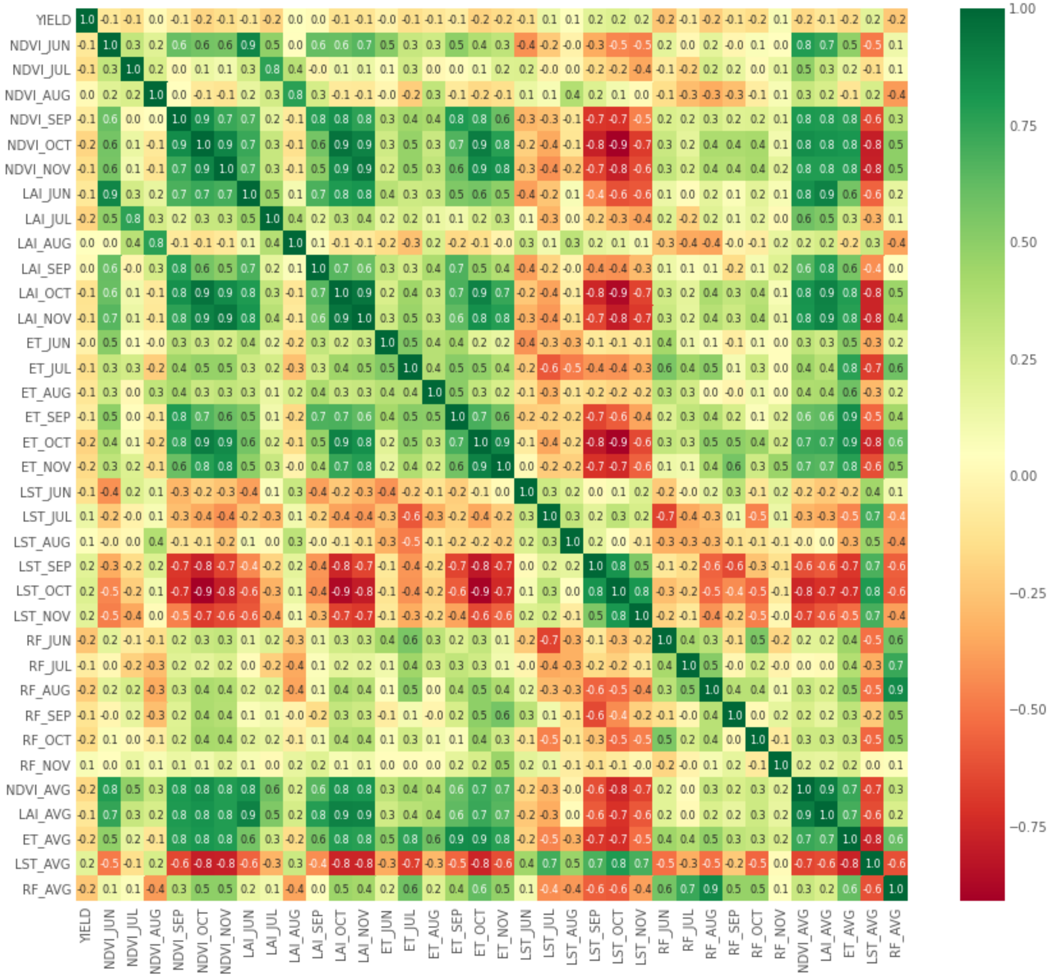
Feature	Month	Difference									Score
NDVI	JUN	-1	0	1	1	-1	-1	-1	-1	1	-2
	JUL	0	0	1	-1	-1	-1	1	1	-1	-1
	AUG	1	0	0	1	-1	-1	-1	1	-1	-1
	SEP	-1	0	1	1	1	1	-1	-1	1	2
	OCT	0	0	-1	-1	1	1	0	0	-1	-1
	NOV	1	0	-1	-1	0	1	0	1	-1	0
	AVG	1	0	0	-1	-1	-1	-1	1	-1	-3
LAI	JUN	-1	1	1	1	-1	-1	-1	1	0	0
	JUL	-1	-1	-1	-1	-1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



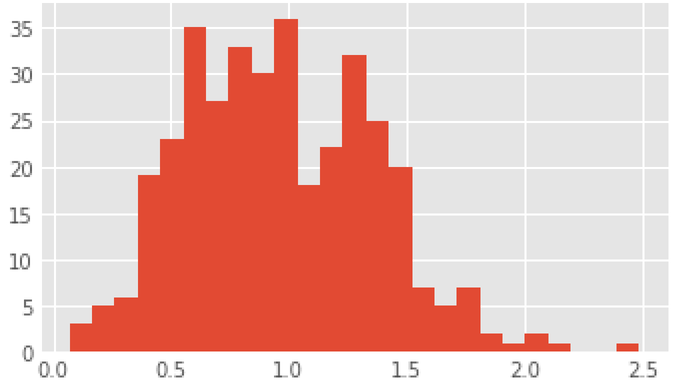
- ❑ We received a similar results comparing with previous table.
- ❑ Also we have checked statistical tests (ANOVA, Mann-Whitney, T-test) for comparing mean values between two or more groups and can remove some features.
- ❑For our simple model we selected the most useful features: *LST_OCT, RF_JUN, LAI_JUL* (these features are not correlated with each other and have the highest correlation with the target) and built simple linear models to predict the target value. 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	0,3264	0,2973	0,307	0,2988
2	0,3276	0,323	0,3181	0,3276
3	0,3147	0,2923	0,2916	0,2858
4	0,3444	0,3278	0,3157	0,3158
5	0,3061	0,2825	0,2781	0,271
AVERAGE	0,3238	0,3046	0,3021	0,2998
STD	0,0145	0,0198	0,0170	0,0227

Target distribution:

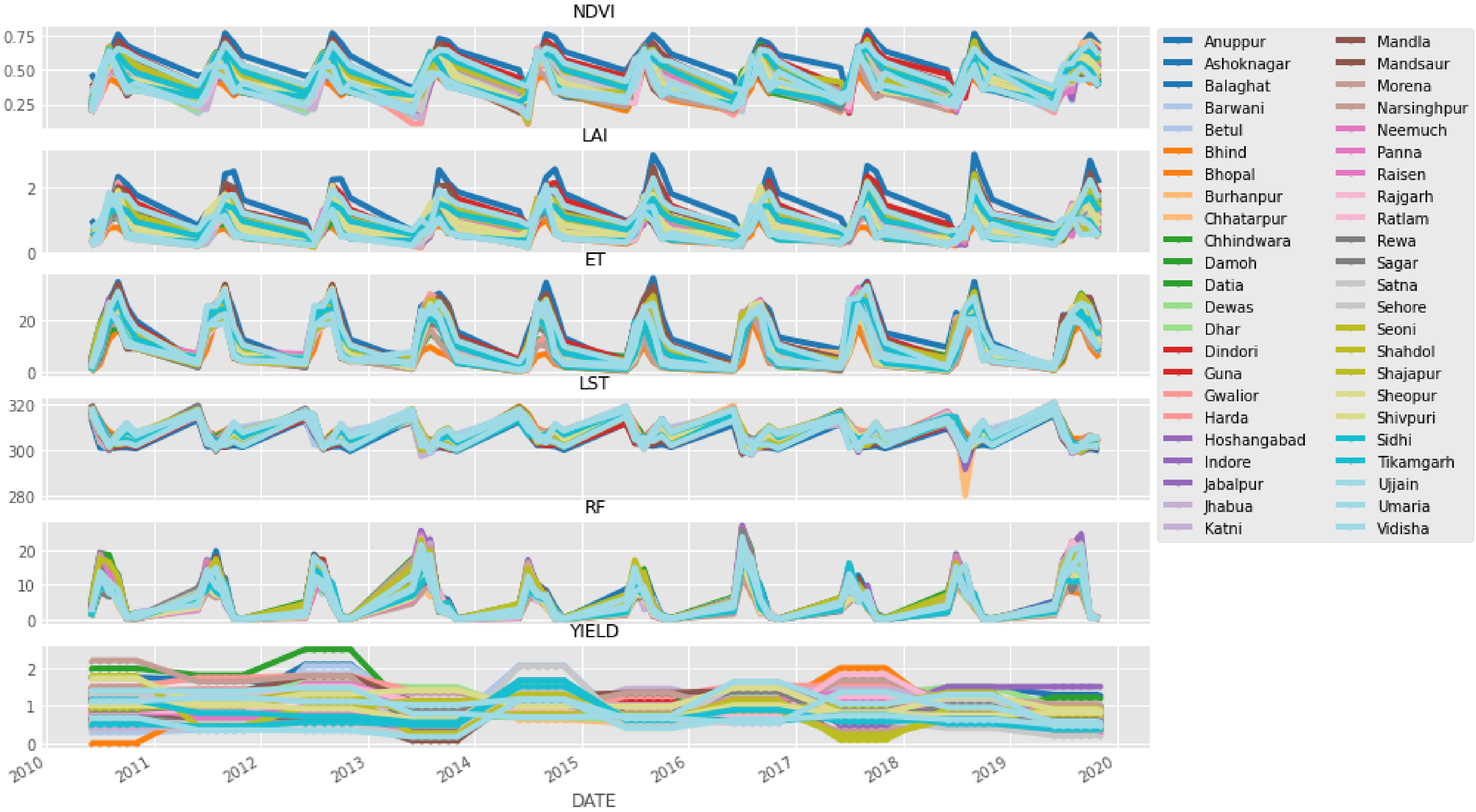


We can see that the most accurate method here is Random Forest model. Our simple method is also have not bad quality and the lowest standard deviation due to simplicity.
By looking at the target distribution plot we see that mean and median value of target is about 1. Hence we have average percentage error of our predictions about 30%.
Simple method have higher bias but lower variance. So we recommend to use it to get predictions for the future Yields. The average formula will be the following:

YIELD = 0.009*LST_OCT - 0.011*RF_JUN - 0.125*LAI_JUL - 1.818

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

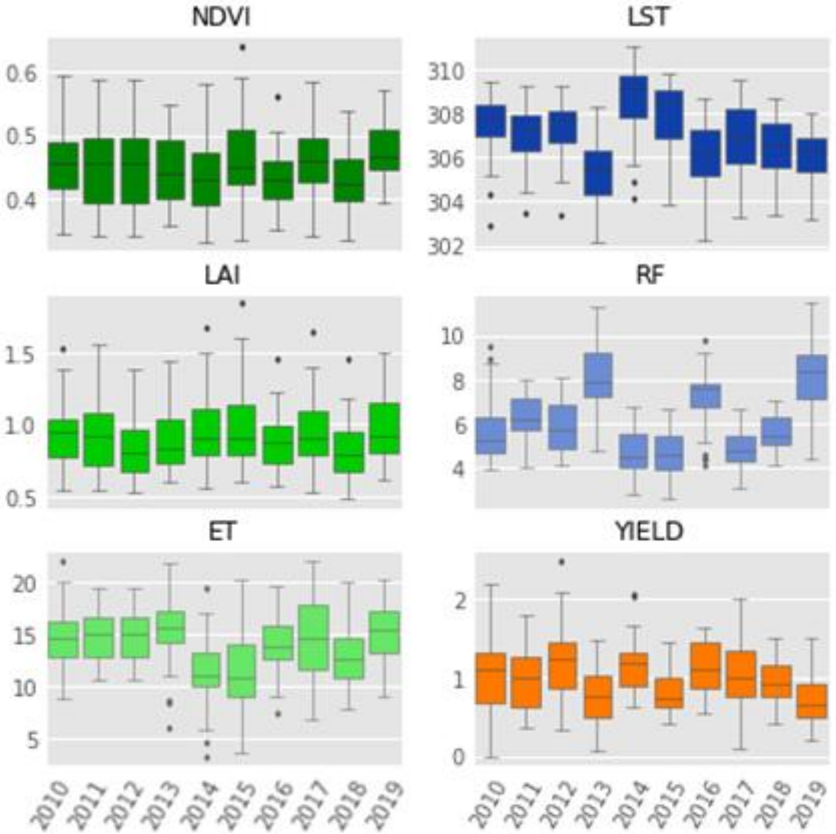
Time series districts plots for every feature



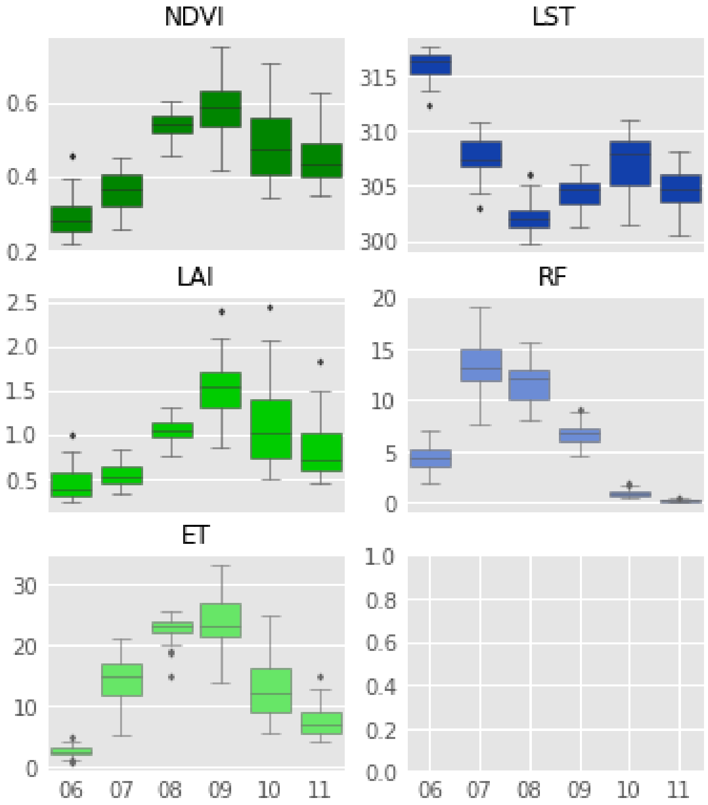
On the picture above we can see all points from dataset. We can see that all districts have similar patterns of feature values. On the next slide we will build separate distributions by years and months and show it on the box-plots and it will be easier to get some ideas from that plots.

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, for example, 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.

Time series parameters

Input datasets:

1. Monthly dataset (6 months for each year) of all features split by different districts.
2. Annual dataset of Yield (target) split by different districts.

Validation set: the last two years: 2018-2019 (for the first dataset - 12 periods, for the second - 2)
Training set: the first 8 years: 2010-2017.

Quality metric: mean absolute error

Models:

- Simple model – last known value,
- Moving averages (1 to 6 - window size),
- Exponential smoothing,
- Prophet model (https://facebook.github.io/prophet/)

Pipeline:

1. To develop time series model that gives the best accuracy on the validation set for ***Features*** predictions.
2. To build **Yield** predictions on validation set based on predicted feature values. We can use here Random Forest model or just simple method described before.
3. To develop time series model that gives the best **Yield** predictions for validation set.
4. Make combination of predictions (2) and (3) as simple average or use both predictions to undestand the range.

*More details you can find in the following script: *2_Time_Series.ipynb*

Features and Yield time series models results

Time series models results:

Method	NDVI	LAI	ET	LST	RF	YIELD
Last value	0,108	0,389	8,972	4,250	6,751	0,422
Moving averages(1)	0,108	0,389	8,972	4,250	6,751	0,422
Moving averages(2)	0,108	0,398	8,161	4,223	6,627	0,348
Moving averages(3)	0,110	0,424	7,374	4,222	6,077	0,309
Moving averages(4)	0,115	0,452	7,458	4,278	5,793	0,285
Moving averages(5)	0,112	0,442	7,359	4,279	5,575	0,289
Moving averages(6)	0,103	0,402	6,731	4,122	5,296	0,268
Exponential smoothing	0,058	0,195	2,830	1,996	2,074	0,400
Prophet	0,066	0,233	3,740	3,165	3,020	0,431

❑ We can see that Exponential Smoothing is the most accurate method for all features. We have used the following parameters for training: ExponentialSmoothing(x, trend="add", seasonal="add", seasonal_periods=6, damped=True)

❑ For Yield prediction moving averages is the best method.

*More details you can find in the following script: *2_Time_Series.ipynb*

Recommendations & Ideas for future work

1. We developed models that can predict future features values and found that Exponential Smoothing are the best methods here.
2. Then these predicted values could be used to predict future Yield values by using simple linear formula of the most useful selected features: ***YIELD = 0.009*LST_OCT - 0.011*RF_JUN - 0.125*LAI_JUL - 1.818*** or by applying Random Forest black-box model.
3. We also tried methods how two predict future Yield values based on their history and found that moving averages is the best method here. These predictions could be considered as the alternative forecasts and compared with the main predictions.

The ideas how to improve results:

1. For time series predictions try to train more methods.
2. Try to extend our initial dataset with new data. For example, you can find datasets about **salaries, road and markets infrastructure and Nitrogen consumption** that stored in the */data/export directory on Github* and have been downloaded from here: http://data.icrisat.org/dld/src/irrigation.html. Possibly there will be some insights from there but unfortunately we didn’t have time to conduct such research.