[SDG 2: Zero Hunger] Crop Yield Estimation and Prediction

Shahira Abousamra
110873138

Gaurav Ahuja
111500318

Mohit Goel
111447365

Abstract

In this project we use satellite sensory imagery to estimate crop yield in a year. We mainly use sensory images representing vegetation index, rain fall, and temperature to estimate the crop yield. We experiment with various models: Linear regression, Random Forest, and SVM, building a crop-specific model and use it to estimate the yield for any country. The report shows our experiments and results, and how well each model performs.

1 Introduction

Our motivation is to develop a reliable, scalable and less -expensive technique for predicting crop yield. Today as the situations stands there is massive amount of crops data available with different organizations and governments but no better way of getting and exploring it. As crop production and yield indicates the state of agriculture, able to monitor and forecast it, play a significant role in ending hunger. With the advancements in big data and computer vision technologies and availability of public remote sensing satellite images, we are doing away with the need to collect the data manually for estimating crop yield.

2 Sustainable Development Goal and Background

Knowing about climatic factors affecting the crop yield and using those to come up with a accurate forecast is central to addressing the food security challenges. It can help us attain United Nations Sustainable Goal 2, 'No Hunger', which aims at ending hunger,food insecurity and malnutrition for all. The need to end hunger become more prominent now by the fact that still 793 million people are undernourished globally. By taking into account different climatic factors we could mitigate the effect of climatic changes over crop yield by taking early and timely steps like : 1 Increasing agricultural investments 2 Introducing Government policies in dealing with low-crop yield 3 Knowing about the countries food reserve and if it seems low than seeking food-aid from high food reserves region.

Previous work such as (Gandhi et al., 2016), (Sabini et al., 2017), (Fan et al., 2015) mainly target yield prediction for a specific country or region. The main conclusion is that deep learning is the best method for yield prediction, as it is now for so many prediction tasks. In (Sabini et al., 2017) They use land classification maps to get crop areas, similar to our method. In general these models are built for countries with wealth of detailed information such as USA, China, and India. We target building a world level model for each crop. This can be useful since city or county level information is hard to get in many areas.

3 Data

The crop yield and climatic indicator datasets used in the project are described below.

3.1 Ground Truth

Food and Agricultural Organization of the United Nation (FAO) provides the ground truth crop yield data set for over 245 countries and territories (FAO,). It has data starting from 1961 to 2013. The dataset includes yields for 187 crops. We chose to work with Maize, Barley and Apples.

3.2 Climatic indicators

NASA Earth Observations (NEO) collects and aggregates satellite data about various physical and climatic indicators like Temperature, Rainfall, Vegetation Index, Solar Insolation at various time granularities like weekly, biweekly, monthly etc (NEO,). We used Vegetation Index, Rainfall,

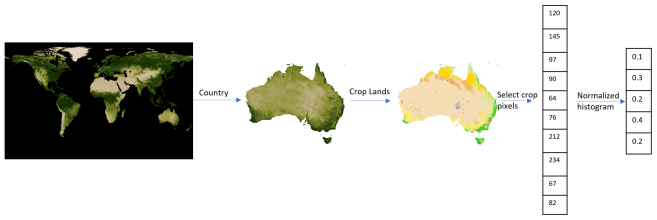


Figure 1: Feature extraction flow

Land surface Temperature (day and night) and Land Surface Temperature Anomaly data for the years 2001 to 2013 to predict the crop yield. The frequency of availability of this data and size is shown in 1.

4 Methods

Our goal is to use global satellite imagery providing climatic data to predict crop yield.

4.1 Feature Extraction

Mapping Pixels to Countries:
Our input features are images representing the world map, where each pixel represents a feature value at the corresponding longitude and latitude. To get country-specific features we need to map each pixel to the country, it belongs to. We used the tool `https://github.com/che0/countries` which relies on GDAL - Geospatial Data Abstraction Library `http://www.gdal.org/` to get the country corresponding to longitude and latitude for a pixel using the shape files 'TM WORLD BORDERS 0.3' `http://thematicmapping.org/downloads/world_borders.php`. The result is a matrix corresponding to the world map where each entry had the UNI code for the country. `https://en.wikipedia.org/wiki/ISO_3166-1_numeric`.

After determining the country to pixel mapping we perform histogram normalization of pixel intensities over those areas which are categorized as potential croplands. The potential croplands are determined using Land Cover Classification imagery dataset, which categorizes different parts of the world under 'Croplands','Wetlands','Forests','Urban Area','Ocean' etc. This ensures that we are considering relevant areas and features for small and big countries come to the same scale. The feature extraction step is described in 1.

Land surface temperature (day and night) and temperature anomaly (day and night) are provided every 8 days. It does not occur exactly weekly. We processed this data to use at biweekly and monthly time scale. From the date of observation, week number of the year was computed. An year consists of 26 bi-weeks. Since this data is coming once every 8 days, it may happen that there are two observations for some bi-weeks, and only one observation for other. For the bi-weeks which have 2 observations, these were averaged to get the representative data and for those bi-weeks which have only 1 observation, that observation was used.

4.2 Standardize Response (Yield values)

Crop yields are standardized to have 0 mean and unit variance. Since the features are now already on the same scale, they are not standardized.

4.3 Spark

We use Spark as data pipeline for computing features from satellite imagery. Features for Vegetation Index, Rainfall, Land Surface Temperature (day and night), Land Surface Temperature Anomaly (day and night), are joined using country and year as key with FAO dataset, to generate our final feature vector. After feature vector formation, we use Spark MLlib for the prediction task.

4.4 Modeling Techniques

We are modeling crop yield using Linear Regression (Plain, Ridge and Lasso), Support Vector Regression (SVR) and Random Forest models. Specifically, We use Spark MLlib implementations of Linear Regression and Random Forest. For SVM we are using Scikit-learn. Since the time overlap of available features and labels is from 2001-2013, the models are trained on data from 2001 to 2010 and are evaluated on data from 2011 to 2013. We create models for 3 crops: Maize, Barley, and apple which has 2150 (train:1649), 1338 (train:1026), 1143 (train:951) observations respectively in the model period, of which 1649, 1026, 951 are training data respectively.

5 Experiments and Results

5.1 Linear Regression

Linear regression is the first technique we experimented with. We show here how our experiments developed, where each step builds upon conclusion from the previous one. We first run our experiments for crop Maize and in the end we create models for 2 other crops: Barley and Apples.
1. First Attempts:

Features	Average Granularity	Units	Range	Time Period	Resolution	Size	Source
Vegetation Index (V)	Weekly	N/A	-0.1,0.9	2000-2013	3600*1800	444.1 M	NEO
Land Surface Temperature Day (TD)	Weekly	Degree celsius	-25.0,45.0	2000-2013	3600*1800	874.7 M	NEO
Land Surface Temperature Night (TN)	Weekly	Degree celsius	-25.0,45.0	2000-2013	3600*1800	856.6 M	NEO
Land Cover Classification	Yearly	N/A		2000-2011	3600*1800	3.45 MB	NEO
Land Surface Temperature anomaly Night (TAN)	Weekly	Degree celsius	-12,12	2000-2013	3600*1800	1.2 G	NEO
Land Surface Temperature anomaly Day (TAD)	Weekly	Degree celsius	-12,12	2000-2013	3600*1800	1.1 G	NEO
Rainfall (R)	Monthly	Millimeters	1.0,2000.0	2000-2013	1440*720	203.5 M	NEO

Table 1: Features

We started off with Biweekly data, with 10 bins histograms, using the yearly data to estimate the yield in that year. The model features were simply the histograms from the sensory images. We experiment with standard linear regression, ridge and lasso regression. Doing a manual grid search over the model parameters and using different sets of features over Vegetation index, temperature day/night, rainfall, and temperature day/night anomaly. The best result was achieved by standard linear regression [SGD step = 0.9] using [vegetation index, rainfall, and temperature Day/Night]. We got a relatively high mean squared error (MSE) given that our values are standardized [MSE for test data = 0.692, MSE for train data = 0.305]. Figure 2a shows the regression results over the test data after re-scaling, where the x-axis is the actual yield and the y-axis is the estimated yield and goal is points on y=x line.

2. Removing Outliers:

Looking at the data and testing some individual countries we notice that countries that have their yield values within some consistent range per the country (such as India or Egypt) give better models, while countries that have jumps in their yield data for example (Algeria) give high mean squared error. So our next attempt was to remove outliers from the training data. We do this by excluding training observations where the yield values are outside the range $[\mu - 2\sigma, \mu + 2\sigma]$ where the mean and standard deviation are over the entire training data, which represents getting rid of the 0.01% outside the normal range. Doing the manual grid search we get best result with Ridge Regression [penalty=0.001, step=0.85], [MSE Test = 0.792, MSE Train = 0.08]. The test MSE increased but plotting the results, figure 2b, it actually looks better with values less scattered and closer to the line y=x .

3. Using Auto-Regression:

We next experiment with auto-regression. Using the value for the previous year as part of the features, Using [penalty=0.001, step=0.85], the results are much better. Without removing outliers, we get [MSE test = 0.172, MSE train = 0.05], figure 2c and with outliers removed we get [MSE test = 0.177, MSE train = 0.03], figure 2d. Again the figure for removing outliers looks better, i.e. points closer to the x=y line, while the test MSE is slightly higher. In both cases when removing outliers, the train MSE decreases. For the rest of the experiments we continue with auto-regression and removing outliers.

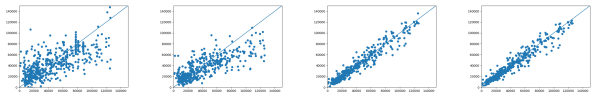
4. Experiment with histogram size:

Since the number of observations for a single crop is in the order of 1000 we do not want to increase the number of features we are using beyond the necessary. We compare the results when using histograms with 10 and 16 bins. If there is significant improvement with 16 we shall experiment with 32. We do the experiment over each feature individually to identify the suitable histogram for each. The feature size = $10 \times 26 = 260$ With 10 bins and $16 \times 26 = 416$ with 16 bins per feature. Table 2 shows the best result achieved by each. It shows very slight improvement for all of them. Based on that we decide to work with 10 bin histograms.

The mean squared error in general is very close, that it is hard to make any conclusion based on the results. While the median squared error has more relative variation. Based on that, We can conclude that the temperature anomaly day and night features have the least contribution to predictions which makes sense.

5. Varying features frequency (Bi-weekly vs Monthly):

We next experiment with varying the frequency of our input. We use all the features together in our model. We model with averaged bi-weekly data and averaged monthly (or more accurately averaged 4-week) data. There are 52 weeks in a year and we are using 6 features with 10 bin histograms, so the feature size is $26 \times 10 \times 6 = 1560$ and $13 \times 10 \times 6 = 780$. Table 3 shows the best result obtained in each case. Again we notice that the relative difference is more accent in the median square error rather than the mean square error with monthly data getting better results. Figure 3 shows yield values predicted by both models against the ground truth.



(a) Plain (b) No Outliers (c) Auto Reg (d) No Outliers

Figure 2: Linear Regression

Feat.	Bins	Model Parameters	Mean SE (Test)	Median SE (Test)	Mean SE (Train)	Median SE (Train)
R	10	reg=l1_p=0.01_step=0.95	0.17196	0.00352	0.02993	0.00300
	16	reg=None_p=0_step=1.0	0.17007	0.00353	0.02862	0.00330
V	10	reg=l1_p=0.001_step=1.0	0.16927	0.00423	0.02883	0.00386
	16	reg=l1_p=0.001_step=1.0	0.16985	0.00395	0.02804	0.00379
TD	10	reg=l1_p=0.01_step=0.7	0.17182	0.00421	0.02965	0.00366
	16	reg=l1_p=0.01_step=0.55	0.17173	0.00415	0.02990	0.00407
TN	10	reg=l1_p=0.01_step=1.0	0.17239	0.00374	0.02964	0.00333
	16	reg=l1_p=0.001_step=1.0	0.17169	0.00453	0.02734	0.00455
TAD	10	reg=l1_p=0.01_step=0.6	0.17097	0.00419	0.02933	0.00408
	16	reg=l1_p=0.01_step=0.45	0.17110	0.00468	0.02982	0.00440
TAN	10	reg=l2_p=0.01_step=1.0	0.17064	0.00761	0.02893	0.00557
	16	reg=l2_p=0.01_step=1.0	0.17116	0.00759	0.02929	0.00543

Table 2: Linear regression with varying histogram bins

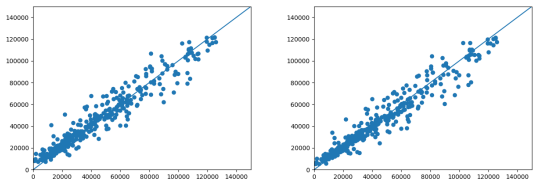
6. Varying period of input and predicting future yield:
We experiment with varying the period of the input data, instead of just one year and estimating the yield for that same year, we add 2 new configurations: 6 months from January to June and 12 months from March of previous year to February of the year in an attempt to predict the future yield.
7. Create models for different crops:
Previous experiments all modeled maize crop. We model other crops; specifically: barley and apples to show the how it works with different crop types that have different growing conditions and number of observations available. Table 4 shows the mean and median squared errors for these models. We see the error is lower for third configuration.

5.2 Random Forest

Random Forests are also commonly used for regression. We experimented with subsets of features and depth and number of trees in the forest. Each feature subset, depth and number of tree configuration was evaluated over test set and the results are shown in Table 5. The errors of each country are also visualized on the world map. Figures 8 to 13 show the results for various crops and time durations.

Frequency	Model Parameters	Mean SE (Test)	Median SE (Test)	Mean SE (Train)	Median SE (Train)
BiWeekly	regularization=l1_penalty=0.01_step=0.5	0.17621	0.00700	0.02910	0.00567
Monthly	regularization=l1_penalty=0.01_step=0.95	0.17182	0.00374	0.02917	0.00359

Table 3: Linear Regression - Varying Frequency



(a) Bi-weekly (b) Monthly

Figure 3: Linear Regression - Varying Frequency

(a) 2011 (b) 2012 (c) 2013

Figure 4: Regression - Apple 1 Year

(a) 2011 (b) 2012 (c) 2013

Figure 5: Regression - Apple 6months

(a) 2011 (b) 2012 (c) 2013

Figure 6: Regression - Barley 1 Year

(a) 2011 (b) 2012 (c) 2013

Figure 7: Regression - Barley 6months

Crop	Input Period	Model Parameters	Mean SE (Test)	Median SE (Test)	Mean SE (Train)	Median SE (Train)
Maize	P1 (Jan-Dec)	regularization=l1_penalty=0.01_step=0.95	0.02917	0.17182	0.00374	0.00359
	P2 (Jan-Jun)	regularization=l1_penalty=0.01_step=1.0	0.02945	0.17280	0.00339	0.00372
	P3 (Mar-Feb)	regularization=l1_penalty=0.01_step=0.95	0.02908	0.17087	0.00347	0.00364
Barley	P1 (Jan-Dec)	regularization=l2_penalty=0.01_step=0.6	0.05818	0.08911	0.01660	0.02154
	P2 (Jan-Jun)	regularization=l2_penalty=0.01_step=0.75	0.06312	0.08892	0.01607	0.01971
	P3 (Mar-Feb)	regularization=l1_penalty=0.001_step=0.55	0.05822	0.08435	0.01612	0.01813
Apple	P1 (Jan-Dec)	regularization=l1_penalty=0.001_step=1.0	0.03013	0.16109	0.00525	0.00738
	P2 (Jan-Jun)	regularization=l1_penalty=0.01_step=0.95	0.03619	0.16451	0.00695	0.00816
	P3 (Mar-Feb)	regularization=l1_penalty=0.01_step=0.95	0.03515	0.16576	0.00582	0.00593

Table 4: Linear Regression - Vary Input Period and Crop

Model Parameters	Crop	FW	Freq	Bins	Period	MSE Train	MSE Test	Median SE Train	Median SE Test
T=50_D=13	Maize	V_TR_TD_TN	Bi	16	Year	0.004	0.421	0.001	0.005
T=30_D=13	Maize	V_TR_TD	Bi	10	6Mon	0.005	0.405	0.001	0.005
T=50_D=13	Maize	V_TR_TD	Bi	10	6Mon	0.005	0.426	0.001	0.005
T=50_D=13	Apples	V_TR_TD	Bi	10	6Mon	0.006	0.315	0.001	0.010
T=30_D=7	Apples	V_TR_TD_TN	Bi	10	6Mon	0.012	0.320	0.004	0.011
T=30_D=13	Apples	V_TR_TAD_TAN	Bi	10	6Mon	0.007	0.315	0.002	0.012
T=50_D=13	Barley	V_TR_TD	Bi	10	Year	0.010	0.175	0.002	0.022
T=30_D=7	Barley	V_TR_TD	Bi	10	Year	0.020	0.170	0.006	0.023
T=50_D=13	Barley	V_TR_TD	Bi	10	6Mon	0.009	0.183	0.003	0.024

Table 5: Random Forest:Vary Input Period and Crop

In these visualizations, red channel was used to indicate overestimation by the model and blue channel was used to indicate underestimation. If the model's prediction is close to ground truth, then the red or blue channels will have small value, and hence the country would appear dark on the map. White color is used to represent the areas for which crop yield is not available.

5.3 SVM

Support Vector Machines can also be used for regression. SVM can map data to higher dimensional subspace using kernels and can potentially provide better results than linear regression. We experimented with taking subset of features, Linear and Gaussian kernels. Grid search was done over the parameter C for Linear kernel and *gamma* and C parameters for Gaussian kernel to find the model with best validation accuracy. The feature set and model combination which performed best over validation set was evaluated over the test set. The results of these experiments are shown in Table 6. The errors of each country are also visualized on the world map. Figures 14 to 19 show the results for various crops and time durations.

(a) 2011 (b) 2012 (c) 2013

Figure 8: RF:Apple 6months

(a) 2011 (b) 2012 (c) 2013

Figure 9: RF:Apple 1 Year

(a) 2011 (b) 2012 (c) 2013

Figure 10: RF:Barley 6 months

(a) 2011 (b) 2012 (c) 2013

Figure 11: RF:Barley 1 year

(a) 2011 (b) 2012 (c) 2013

Figure 12: RF:Maize 6 months

(a) 2011 (b) 2012 (c) 2013

Figure 13: RF:Maize 1 Year

Model Parameters	Crop	Feat.	Freq.	Bins	Period	MSE Train	MSE Test	MedianSE Train	MedianSE Test
C=1_g=0.01	Apples	V_R_TD_TN	Bi	10	6Months	0.029	0.169	0.007	0.008
C=1_g=0.01	Apples	V_R_TD_TN	Bi	16	6Months	0.028	0.166	0.008	0.008
C=10_g=0.001	Apples	V_R_TD_TN	Bi	16	1Year	0.025	0.161	0.008	0.008
C=10_g=0.01	Barley	V_R_TD_TN	Bi	16	6Months	0.03	0.099	0.01	0.019
C=1_g=0.01	Barley	V_R_TD_TN	Bi	10	1Year	0.04	0.106	0.01	0.02
C=10_g=0.01	Barley	V_R_TD_TN	Bi	16	1Year	0.017	0.094	0.01	0.024
C=10_g=0.001	Maize	V_R_TD_TN	Bi	10	1Year	0.019	0.179	0.005	0.005
C=10_g=0.001	Maize	V_R_TD_TN	Bi	16	1Year	0.018	0.177	0.005	0.005
C=10_g=0.001	Maize	V_R_TAD_TAN	Bi	10	1Year	0.019	0.173	0.005	0.004

Table 6: SVM:Vary Input Period and Crop

(a) 2011 (b) 2012 (c) 2013

Figure 14: SVM-Apple 6months

(a) 2011 (b) 2012 (c) 2013

Figure 15: SVM-Apple 1 Year

(a) 2011 (b) 2012 (c) 2013

Figure 16: SVM-Barley 6months

(a) 2011 (b) 2012 (c) 2013

Figure 17: SVM-Barley 1 Year

(a) 2011 (b) 2012 (c) 2013

Figure 18: SVM-Maize 6months

6 Discussion

The results show that it is possible to create a model for crop prediction using climatic indicators which generalizes well for most of the countries. It is also observed that crop yield in an year can be predicted well by using data for first six months of the year.
For linear regression, we notice that lasso regression usually gives the least error. This coincides with the fact that the number of features is big almost as large and sometimes larger than the number of observations. Auto-regression was important to achieve good results. We tried modeling with single countries and the results varied based on how much outliers it contained. This sort of explains the behavior of the model with so many values and jumps that cannot all be fit by linear regression even after excluding observations outside the 99% region. This agrees with the observation that the difference in the mean squared error among best models in each case is usually small while the median squared error is better in capturing the difference among the models.

For SVM, we observe that RBF kernel gives us better results over Linear kernel. The results with SVM are close to those obtained with linear regression. It suggests that allowing the model to learn non-linear boundary did not improve the results much.

For maize, linear regression outperforms both random forests and SVM. All the three models perform similar on Barley. SVM performs slightly better than linear regression for Apple. They both outperform random forest for Apple.

(a) 2011 (b) 2012 (c) 2013

Figure 19: SVM-Maize 1 Year

It can be said that the best performing model varies from crop to crop. Hence if this method is extended for many crops, all the three models should be tried and the best one should be chosen for every individual crop.

7 Conclusion

Estimating crop yield beforehand prepares the government and public to be better prepared for the future. Our model tries to predict the crop yield taking climatic factors into account. Using combination of different climatic conditions we tried to estimate the effect of them on crop yield and it gets certainly proved that some factor like Rainfall and Vegetation Index are positively co-related with the crop production. Also we train models separately for different crops as weather impacts differently every other crop. One pronounced result is that using features for first half of the year are equally good in predicting crop yield in comparison to the features for the full year.

8 Team Member Contributions

Shahira: Discovering data, Mapping pixels to countries, Combining individual features to generate features, Linear regression experiments, Report writing

Mohit: Discovering data, Getting and making data accessible, Unification of imagery, Processing Land cover classification data, Processing FAO dataset, Feature extracting for monthly data, Linear regression experiments,Developing experiment framework for Random Forest and SVM, Random Forest experiments, Report writing

Gaurav: Finding data, Mapping pixels to countries, Converting images to final features, Linear regression experiments, SVM Regression experiments, Visualizing results, Report writing

References

Wu Fan, Chen Chong, Guo Xiaoling, Yu Hua, and Wang Juyun. 2015. *Prediction of crop yield using big data*. 8th International Symposium on Computational Intelligence and Design.

Food and agricultural organization of the united nations (fao). http://www.fao.org/faostat/en/#data. Data source.

N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy. 2016. *Rice crop yield prediction in India using support vector machines*. July.

Nasa earth observations (neo). neoftp.sci.gsfc.nasa.gov. Data source.

Mark Sabini, Gili Rusak, and Brad Ross. 2017. *Understanding Satellite-Imagery-Based Crop Yield Predictions*. Stanford University.