AI Hackathon 2021

Challenge: Crop Yield Challenge
Team: We are Dense

Imperial College London

Name: Abu-Al Hassan, Danila Kurganov, Saleh Komies and Helen Li

Date: 21/02/2021

Members Background: Mathematics and Electrical Eng Departments

Crop yield estimation prior to harvesting period with an AI approach.

Abstract (extended)

Knowing an estimate of the yield of an agricultural product prior to the harvest period can help governments to introduce policies and farmers to stand a better position in trading with future contracts and prepare supply orders. We further aim to extract meaningful insights the data implicitly carries in with, which is later seen to mimic 'industry knowledge' yield info.

The overwhelming largest crop output in Illinois are corn and soybean. These crops have their planting season in April and harvest season in October and November. We aim to predict the yield of crops for a year in advance of the harvest time. Thus we have trained our model only on data between April-September.

Our model uses EVI [*1] (every 16 days) and 2m temperature[*2] (1st, 15th and 28th of every month) from the start of April until the end of September as well as latitude and longitude readings.

1. Data exploration

In order to start looking at different features in the datasets, we have begun with exploratory data analysis. From this, we have identified some trends and insights into the dataset.

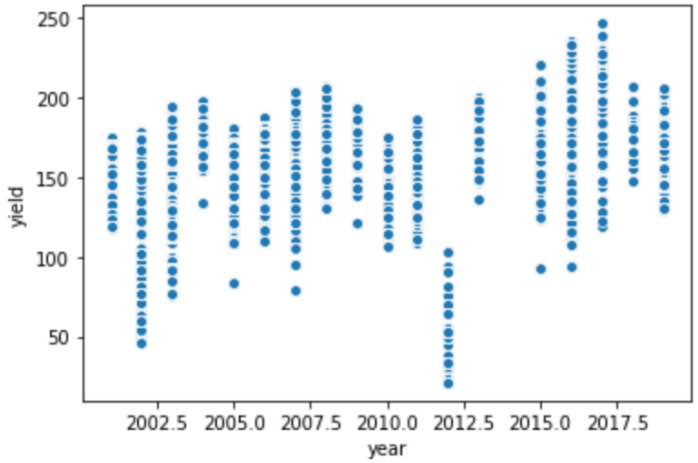


Fig 1. The yield on average seems to be increasing and some years have an abnormally low yield which may be a result of extreme weather such as the biggest dip in this graph that happens during the 2012 Illinois drought.

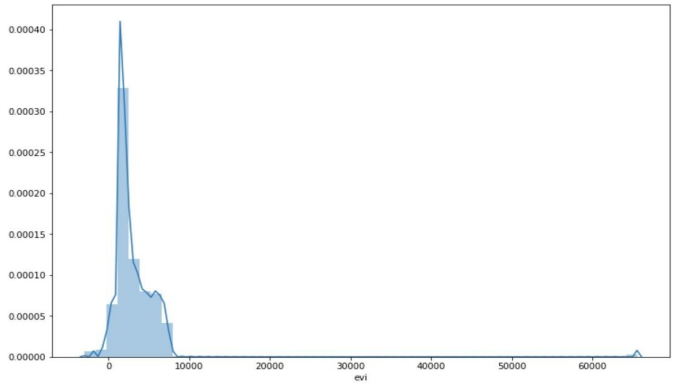


Fig 2. Outliers in the right end of data hence removal of extreme values (EVI>10000 [1])

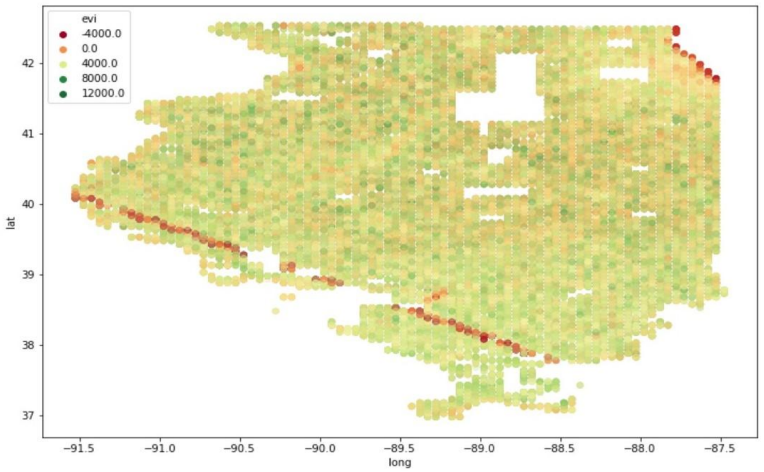


Fig 3. Straight red line of low EVI on the south-west part of Illinois, which we identify as being close to the Mississippi River. Straight-line on the north-east part of Illinois corresponds with the boundary of Lake Michigan. The low EVI here is due to flooding.

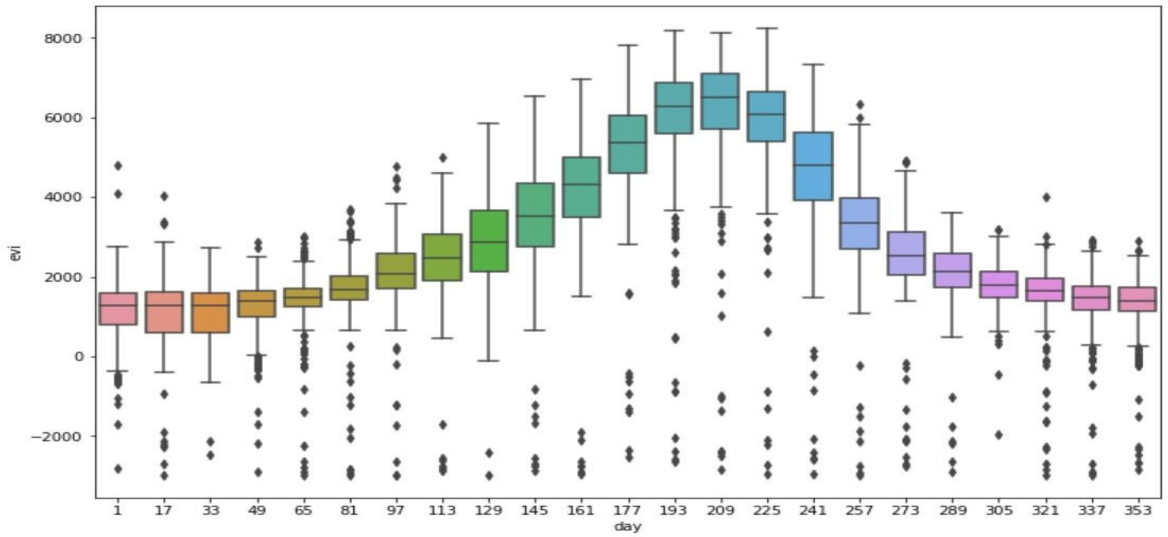


Fig 4. EVI varies at different times of the year. Starts to increase from the planting season in April, peaks in June/July and is very low during the harvest in September and October. The EVI also decreases during August and September, this can be explained due to crops becoming ripe and thus decreasing in their chlorophyll levels. EVI is responsive to chlorophyll and hence decreases.

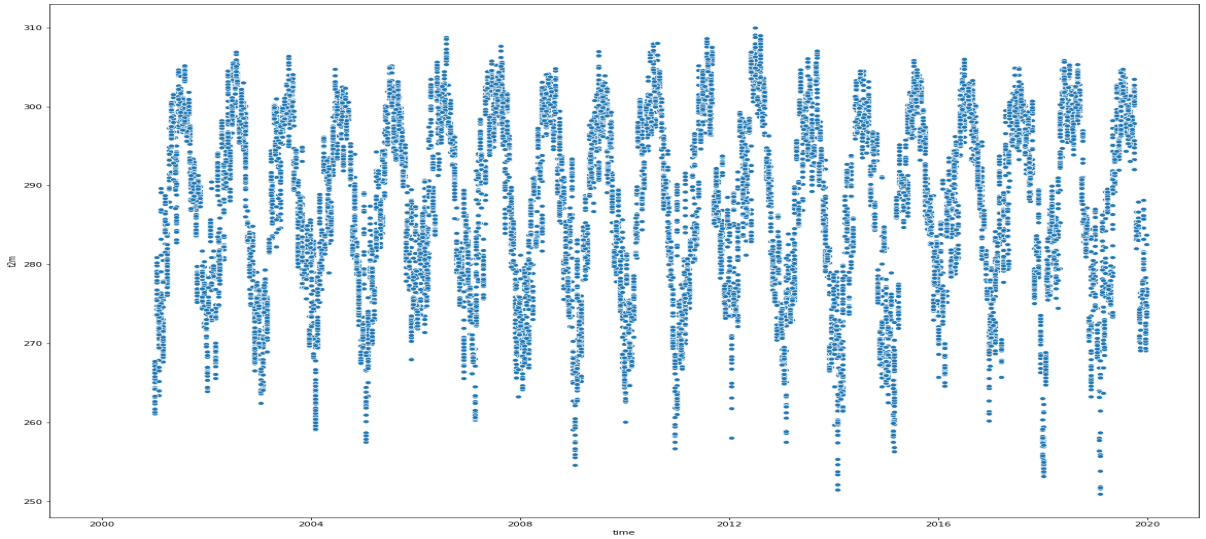


Fig 5. The temperature profile over the years in Illinois stays relatively the same but with some extreme winters with temperatures of -21.3℃ (250K as shown in t2m reading). The summer temperature goes as high as 36℃ (309K).

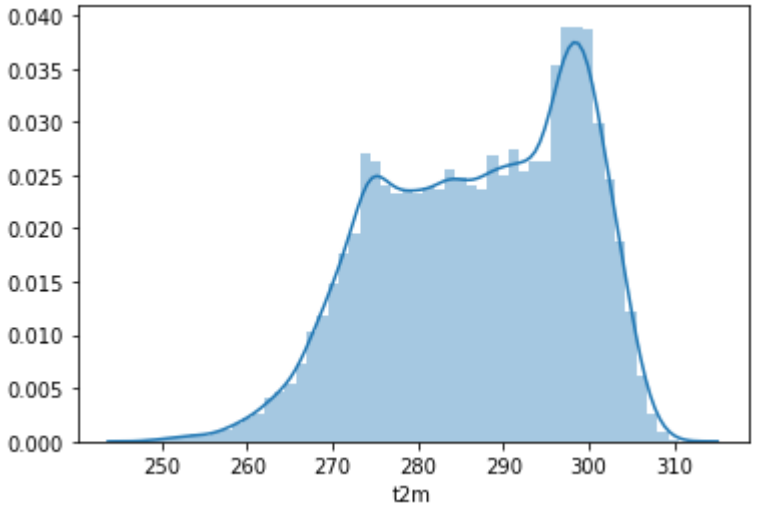


Fig 6. The most common temperature over the last 20 years seems to be around 300k (~26℃). The spread of high temperature is not significant but some of the low temperatures are likely to be due to snowfall that exceeded 80inches in 2013/14 [3]

In order to align both ERA5 and Illinois yield data to EVI_stacked data, we have only used data between 2001-2019 and April-September from each dataset and concatenated on the latitude, longitude and year. Before concatenating we turned the t2m column in ERA5 to many temp_{day of year columns} as this gives the model a sense of time even though we don't have a time column in our final data frame. We then dropped irrelevant columns in our dataframe. Our final dataframe has the following columns:
['long', 'lat', 'evi_97', 'evi_113', 'evi_129', 'evi_145', 'evi_161', 'evi_177', 'evi_193', 'evi_209', 'evi_225', 'evi_241', 'evi_257', 'evi_273', 'yield', 'temp_91', 'temp_105', 'temp_118', 'temp_121', 'temp_135',

'temp_148', 'temp_152', 'temp_166', 'temp_179', 'temp_182', 'temp_196', 'temp_209', 'temp_213', 'temp_227', 'temp_240', 'temp_244', 'temp_258', 'temp_271', 'temp_274']

We create a train test split to our dataframe to test out our models. In order to visualize the enhanced vegetation index using the Illinois-counties.geojson and EVI_stacked which respectively contains the geometries of counties in Illinois and EVI observations for each 16 days in the column. We have included five different random figures for the years (2001, 2008, 2013, 2015 and 2019). We see EVI_Stacked shows major changes in recent years.

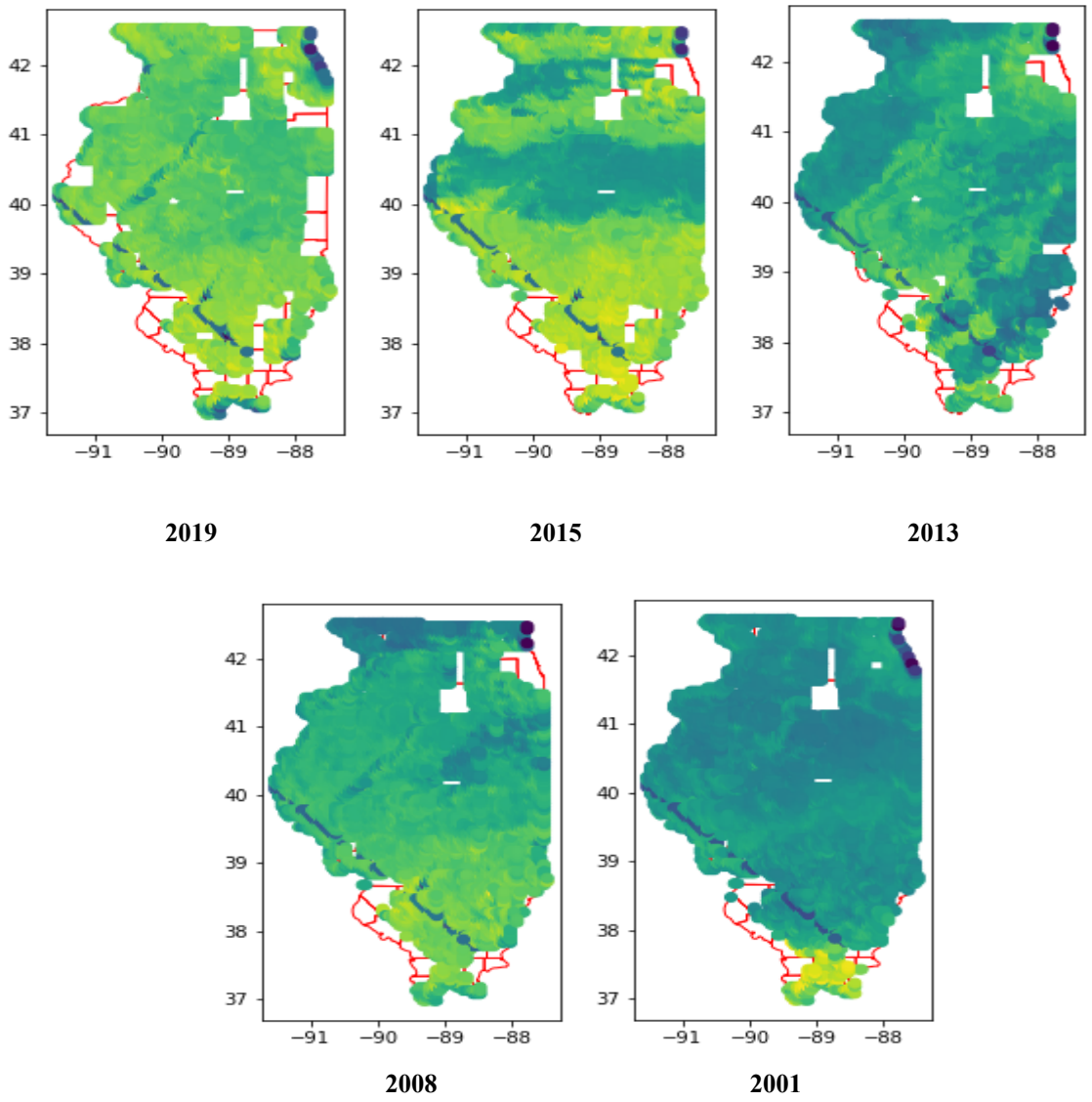


Fig7. EVI stacked data in Illinois for each 16 days for random years 2001 till 2019

1. Methodology

Models

1. Sequential neural network model

The sequential model involves using the output of the previous layer as an input of the current layer. For each of the 3 hidden dense layers, we have used the same number of nodes as the input node (33) and returns a singleton output node in the output layer. Through each of the dense layers we have used ReLU as the activation function, due to 2 reasons:

I. We are constructing a regression model to predict the number of crop yields in Illinois, hence a linear model such as ReLU should be at least used in the output layer of the model

II. Try to avoid the vanishing gradient problem that occurs in multi-layer networks and limit the problem to be train-able on a local CPU

The ReLU function is given by:

$$f(x) = max(0, x)$$

Which is a piecewise linear function that outputs the value x if it is positive, and otherwise 0. For compiling optimisers, we have chosen to use Adam, which has the ability to perform well with noisy/sparse gradients and requires small memory space. After observation of the reduction in validation losses due to the change in the number of epochs, we have fixed the number of epochs to 400, both validation and training loss after each cycle reduces consistently before epoch = 300, but levels off after this point.

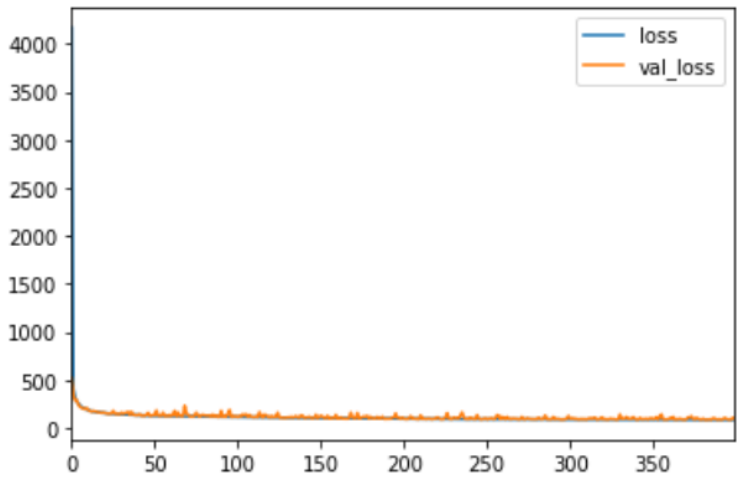


Fig8. Loss and validation loss against the number of epochs, which both significantly declines from 4000 in the first few cycles and stabilized around 83

We have later experimented with adding in dropout layers, call-backs and adjusted the number of nodes in dense layers in an attempt to reduce the validation loss further. However, we did not reach a stage that validation loss is significantly reduced.

2. Decision tree regressor

A Decision Tree Regressor uses predictors that minimizes RMSE at each split of the branches while choosing the root being the best predictor in minimising RMSE of predictions to actual data points. However, this method did not provide us with the best model of the lowest RMSE in our predictions when comparing to other approaches(10.54 and 0.86 respectively). But has enlightened us to continue using a tree-method prediction but randomize the creation of each decision tree, hence a Random Forest Regressor.

3. Random Forest Regressor

A random forest regressor combines multiple tree regressors to generate a final output. The motivation of using a random forest regressor from sklearn is this algorithm runs efficiently on large datasets, which is essential for training models on local CPUs. After using the same model over different numbers of trees (33, 50, 70, 100, 200, 500), we did not see a significant reduction of RMSE in the predictions. Hence we are inclined to use 33 as the number of trees to use in the random forest regressor to reduce CPU usage.

4. Base model architecture based on end-consumer's goals and useful parameters

Random forests were first used to analyse and clean data via feature extraction techniques. This includes examining decision tree branches, OOB error, and finding data-leakages. For model interpretability, low-significant and redundant features were removed. A NN was trained on this cleaned data. Due to similar errors, we ensemble both models which makes the final model. Using training data with no industry-insight input, feature extraction allows us to reverse-engineer some tips used for planting the main crops found in Illinois.

3. Results

1. Sequential neural network model

From this approach we have achieved a relatively accurate prediction in crop yield in October/November using data from April to September. The RMSE and explained variance from this approach is around 9.57 and 0.89 respectively.

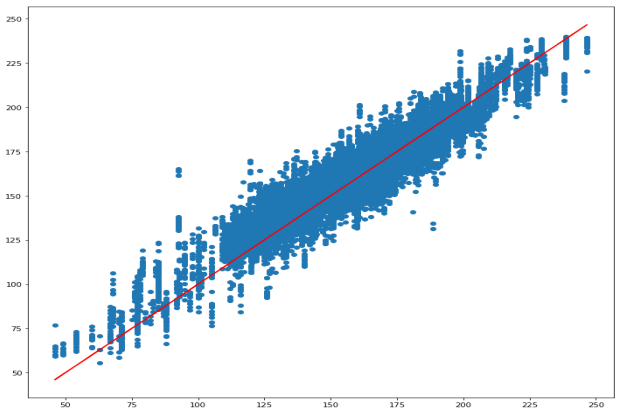


Fig9. Plot of test data (blue scatters) against actual predictions (red line).

This shows that our prediction lies fully inside the range of response variables, hence indicates the suitability of this model.

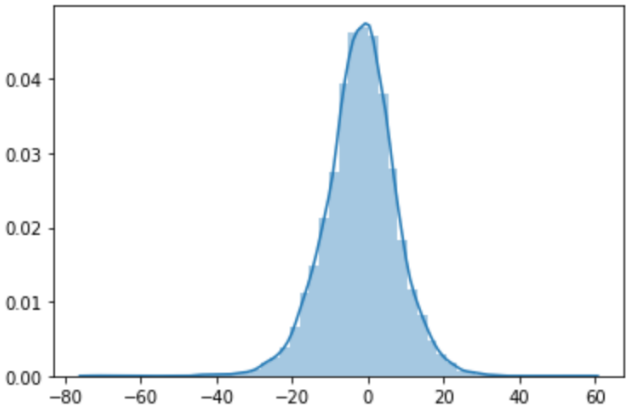


Fig 10. Histogram of errors from neural network approach. This is roughly standard normally distributed.

From the plot we can tell that irreducible errors are normally distributed, which is confirmed with a variance of ~ 0.89, which solidifies the suitability of this model.

2. Decision tree regressor model

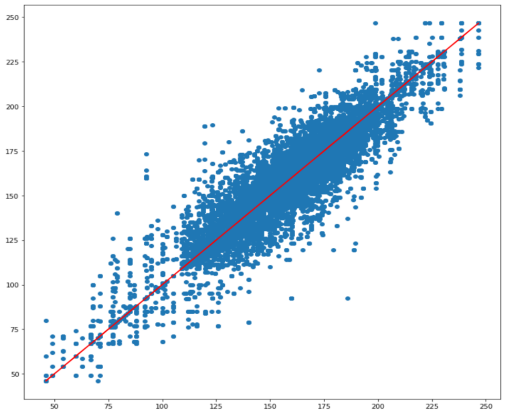


Fig 11. Scatter plot of test response variable values (blue) against predictions from the model (red)

Figure 11 shows a similar fit to the previous neural network approach, which confirmed the suitability of a tree-method model to this challenge. This motivated us to use further tree approaches based on a Decision Tree.

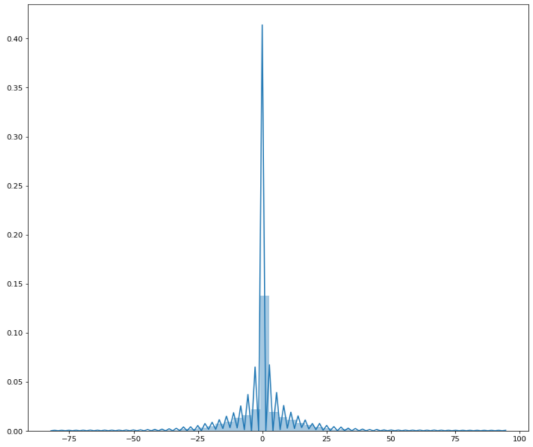


Fig 12. Histogram of errors from decision tree approach. Symmetrical distributed but saw-teeth type.

Our idea of suitability is slightly weakened by the shape of this plot, but since the overall shape is roughly normal centred at 0, combined with a variance ~ 0.86, hence we are motivated to continue with a tree-method approach.

3. Random forest regressor model

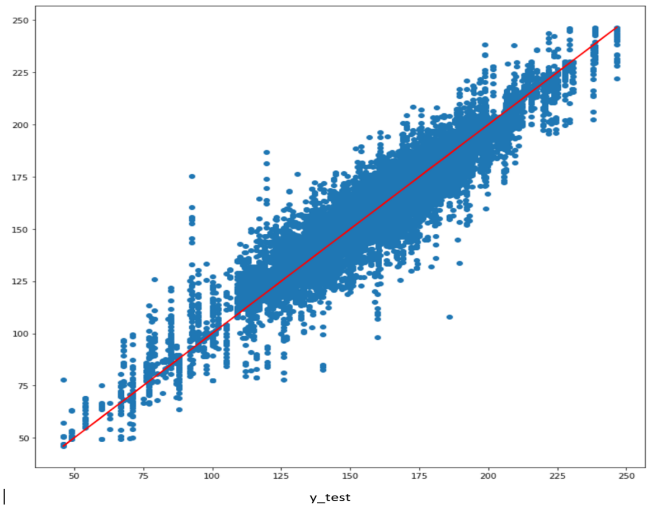


Fig 13. Scatter plot of test response variables (blue) against predictions from random forest regressor model (red)

The graph indicates some relatively large error in low-valued predictions (around 50) but consistent overall. Such that the prediction values all lie within the range of values in testing response variables.

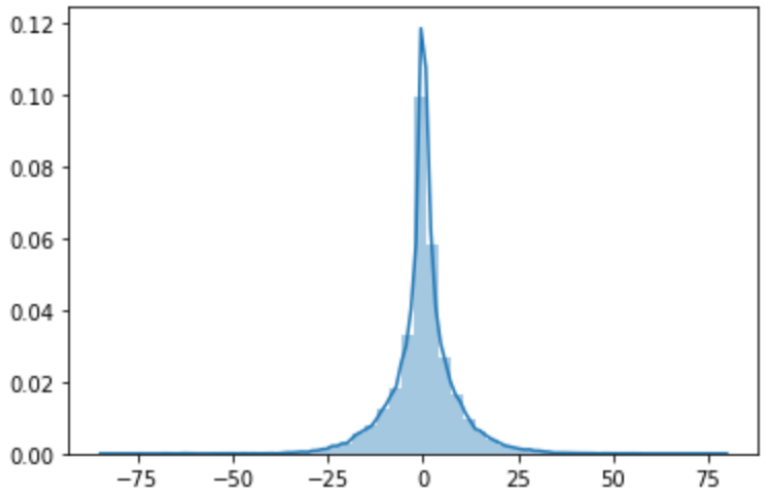


Fig 14. Histogram of errors obtained from random forest regressor

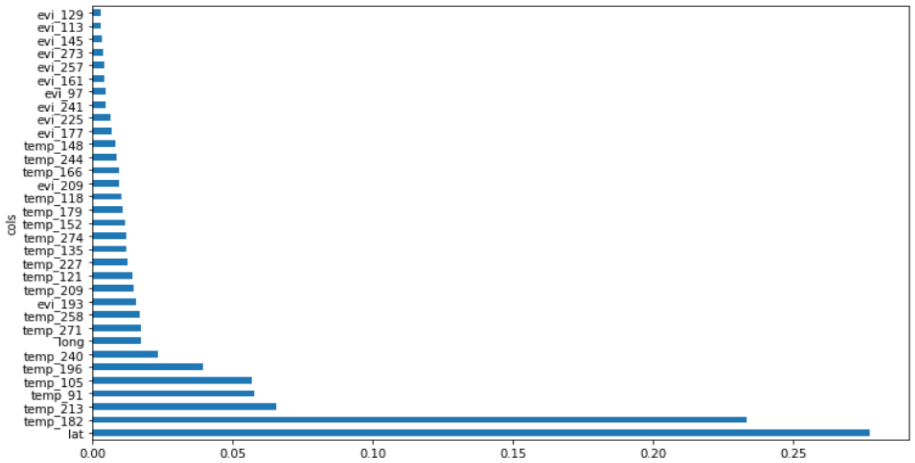


Fig 15. Importance of each of the features used in random forest model

This enables us to highlight that latitude and temperature in the second-half year has a high importance in predicting the final value.

4. Final random forest model

Our Random Forest model and NN gives training and validation RMSE of 7.278, 7.623 and 7.22, 7.76 respectively. The combined model further gave a 5.1497 mean absolute error, and 0.934 explained variance score. As MAE scores ignore outliers better, this value gives good predictive power given the yield range of 200 (246-46).

For model confirmation, the below scatter plot shows relation between prediction and actual values for the validation set. These results heavily mimic Abu-al's data.

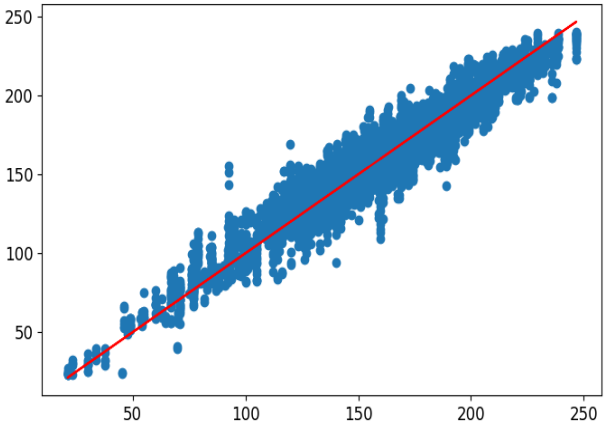


Fig 16. Scatter plot of test response variables (blue) against predictions from random forest regressor model (red)

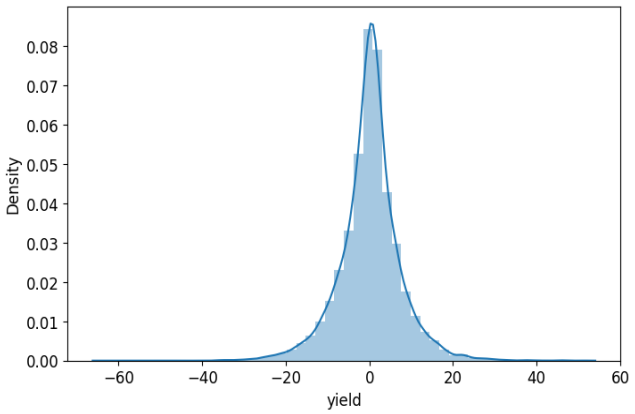


Fig 17. Histogram of errors

We can determine feature which our model finds most important

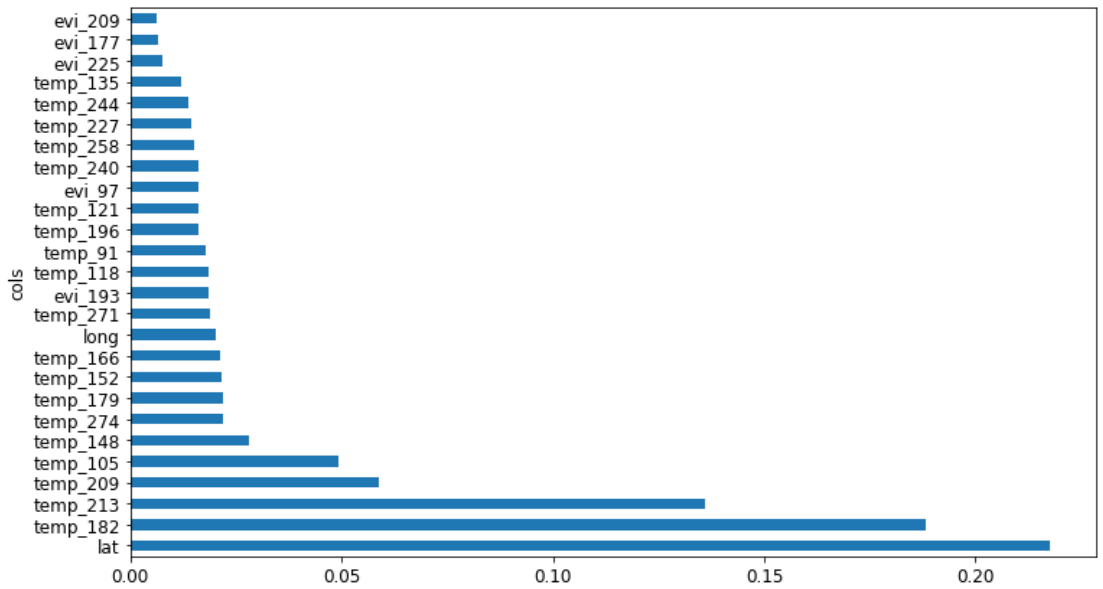


Fig 18. Importance of features used in the random forest model

From this, we can tell that latitude has the largest importance in the predictions generated by the random forest training. Surprisingly, EVI over the 4 months observation window did not provide a strong aid in computing predictions. But the temperature of days around the end of growing period (August time) and middle period (June) has a high importance in projecting a good harvest at the end of year.

The following plot shows how the model's predictions are affected over time as various data points are assessed.

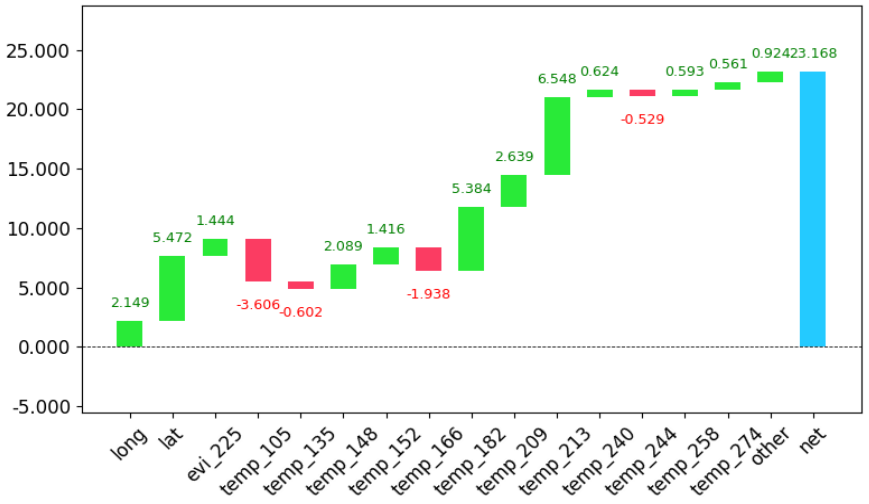


Fig 19. Waterfall graph of the quantified effect of predictors on final prediction

4. Data Analysis with final model

Using our data, we can further confirm plant growing information with partial dependence.

Location vs. Crop Yield

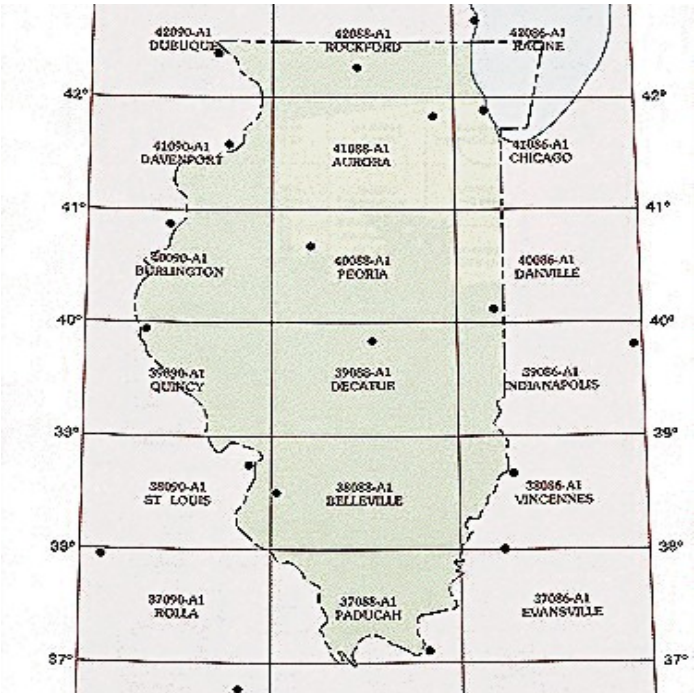


Fig 20. Partial map of latitude and longitude in Illinois

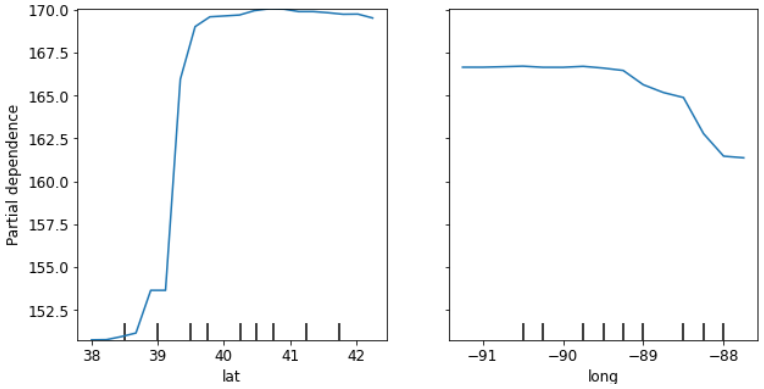


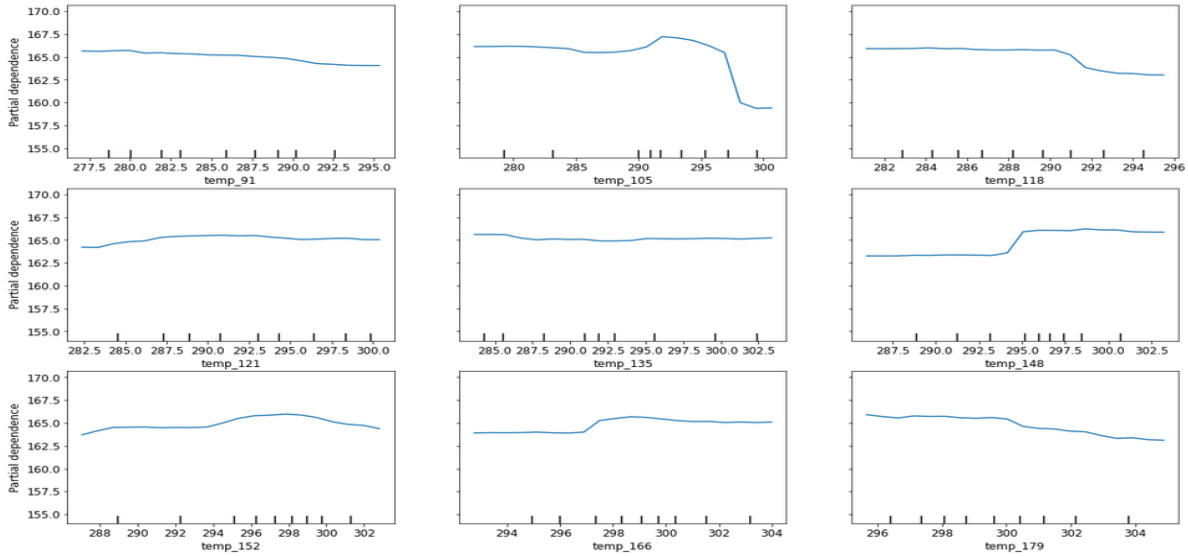
Fig 21. Partial dependencies of predictions on latitude and longitude

We see how, under all other conditions being the same, plant yield does not differ above 40* latitude line, and is not as good under this line. Considering soil-moisture data[4], there is a chance that latitude is simply partly encoding soil-moisture given their observed correlations to crop yield.

Corn is a warm season plant that requires large amounts of water only in the 30 minutes period after saturation [5], too much soil moisture is one of the main factors for reduced corn crop yield for corn, as well as soybean crop yield [6].

Generally speaking, the 'best' place to plant corn is in the north west. This is a subject with high potential variance, given that regions of IL that only have small land-area with certain latitudes and longitudes may simply have few crop-yielding locations, in which case one poorly performing location will affect the entire region more significantly than in other places.

Extreme Temperatures vs. Crop Yield



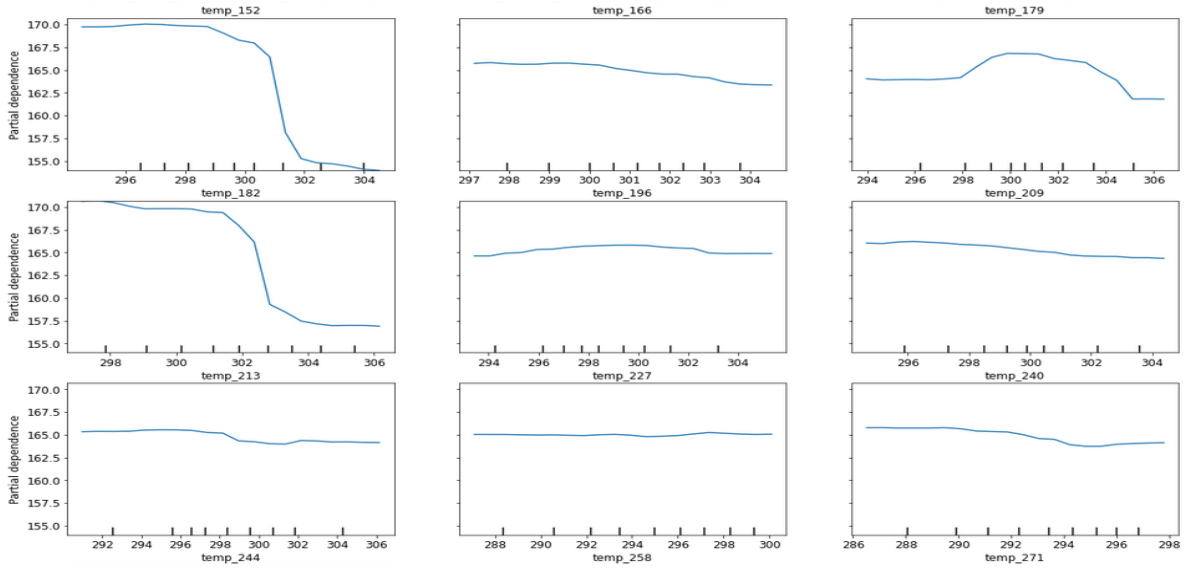


Fig 21. Partial dependencies between temperatures in the observation period and predictions.

The above show the change in yield given certain weather events at a specific day in the year.

Notable observations:

1) Hot weather (>297) on day 105 highly affects crop yield for the whole year. Our model finds day 105 significant. In real life, we note that this is the time at which most plants are being planted [source for illinois plants by mass by planting]. Furthermore, at this time, sources say it is best to plant corn when soil temperatures are at 283K[7], and 292K for soybeans. This 'best practice' advice is mimicked by our model somewhat.
2) During the summer months, extreme hot temperatures decrease yield significantly; I believe this is due to dry-ness of soil, which was previously mentioned to affect crop-yield. It's interesting to note which days were actually significant for this analysis - however given that the model was only trained on a few years of time, and hot temperature events are not localised across illinois, this variance in day significance could be due to low data. In any case, the decreasing yield to temperature trend is similar across a grouped period of time representing june-july, so there is some significance to this prediction and theory.
3) The final trends representing no particular change in yield towards the end of the growing periods could have several interpretations. I From having lived in Illinois, and noticing that the august-september month was more humid and rainy than june-july months, it could be that plants did not dry-out as much due to the extra moisture.

Landsat Enhanced Vegetation Index vs. Yield

We immediately see that any changes to EVI lead to, at most small changes in yield. Instead of considering a 'change in satellite imagery', we interpret EVI change as change in vegetative greenery on yield. A question this might pertain too is "If I look at a crop field and it's somewhat green, how much more yield can I expect if these crops on this day happen to be *very* green?"

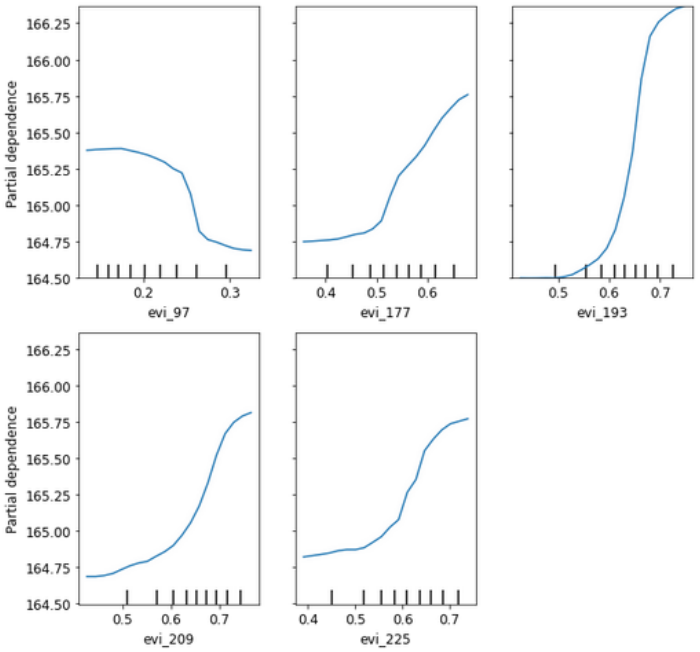


Fig 22. Partial dependencies of EVI's used in the NN

We can ignore the first column data since this is right when seeds are being planted, and it makes no sense to judge vegetative greenness at this point in time as a predictor for yield. All other days, much later in the plant growth cycle, produce plots which very much make sense with our intuition. Just observing sweet-corn data [8], we see that after around 100 days of seeds being planted, in the maturity stage of corn-growth development, the EVI index is quite significant on yield. This can be interpreted as: During the maturity stage of corn, how green it looks has a highly correlated prediction for the yield we achieve.

What we could have done better:

We failed to consider crop rotation data. On the bright side, our models allow for economical deployment due to easy to gather input data. Using publicly available data implementations could have been as beneficial as transfer learning, with more interesting partial pivoting insights. We could have used NN-embeddings to view spatially-correlated data.

5. Conclusion

The data evaluation was open-ended, hence, considerable time was spent to evaluate which data is useful, and what insights would be meaningful to an end-user. We came to the conclusion that crop yield data would be most useful to an end user if the model could predict yield before the time at which crops were harvested. We therefore trained our model so that in hindsight, it can predict crop yield 2 months into the future. We believe this gives croppers ample times for harvesting logistics. Given our Illinois-centric data, our data analysis was based on crop-type information from this state. In production, our model can be deployed to give yield information, updated over time, from the start of planting to the harvesting phase. Model evaluation shows us little need for EVI data apart from

sparse but highly-correlated forecasting in early June. This gives good forecasting in an economical way for the end-user.

Our model is based on a sequential-like ensemble between a NN and Random Forest. This provides robustness for future extreme weather events and yields predictions, as well as minimising errors. For production the ensemble is used.

Interpreting the Random Forest model gave us interesting data insights. We infer IL locations with the best yields, and how variance of temperature on specific days predicts yields. Notable insights, mimicking seed-planting best-practices, were the strong correlation between air temperature at the time when seeds were planted with overall annual yields, and generally stronger yield with higher temperatures (up to drought). One insight, not found in literature, was EVI correlation during specific times with yield. One interpretation is as follows: the greener the crop after 100 days, the better the yield.

Improvements could be done in the model making phase. This includes NN-embeddings, and using more historical data to predict yields. This will encode information such as crop-rotation and other possibly useful information. This would be less useful for the farmer, given their insight, but more useful for end-users unaware of farm-specific data.

6. Glossary

[*1] EVI - Enhanced Vegetation Index, this is a quantified value for the amount of greenness of the vegetation
[*2]2m temperature - Temperature of air measured 2m above ground/water level.

7. Reference

[1] "Landsat Surface Reflectance-Derived Spectral Indices" https://www.usgs.gov/core-science-systems/nli/landsat/landsat-enhanced-vegetation-index?qt-science_support_page_related_con=0#qt-science_support_page_related_con (accessed 21 Feb 2021).

[2] "Chicago - Highest Temperature for Each Year." https://www.currentresults.com/Yearly-Weather/USA/IL/Chicago/extreme-annual-chicago-high-temperature.php (accessed 21 Feb 2021).

[3] "Climate of Illinois." https://en.wikipedia.org/wiki/Climate_of_Illinois#Thunderstorms_and_severe_weather (accessed 21 Feb 2021).

[4] "Water and Atmospheric Resources Monitoring Program (WARM)." [Online]. Available: https://www.isws.illinois.edu/warm/.

[5] "Soil Temperature and Corn Emergence." https://www.pioneer.com/us/agronomy/soil_temp_corn_emergence.html (accessed 20 Feb 2021).

[6] "PLANTING SOYBEANS IN SOGGY SOILS LEADS TO LONG-TERM PROBLEMS." [Online]. Available: https://www.agriculture.com/crops/soybeans/planting-soybeans-in-soggy-soils-leads-to-long-term-problems.

[7] "Corn planting: Consider soil temperature and date." [Online]. Available: https://crops.extension.iastate.edu/encyclopedia/corn-planting-consider-soil-temperature-and-date.

[8] "Back to Vegetables A to Z - Soya beans." https://www.rhs.org.uk/advice/grow-your-own/vegetables/soya-beans (accessed 21 Feb 2021)

[9] "Sweet corn growing." https://www.pinterest.dk/pin/615727198949155 82/ (accessed 20 Feb 2021).

[10] L. Martínez-Ferrer, M. Piles, and G. Camps-Valls, "Crop Yield Estimation and Interpretability With Gaussian Processes," *IEEE Geoscience and Remote Sensing Letters,* pp. 1-5, 2020, doi: 10.1109/LGRS.2020.3016140.

[11] A. Kaneko *et al.*, "Deep Learning For Crop Yield Prediction in Africa," 2019.

[12] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian process for crop yield prediction based on remote sensing data," presented at the Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 2017.

[13] M. Ilse, J. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," presented at the Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, 2018. [Online]. Available: http://proceedings.mlr.press.

[14] A. Viña, A. A. Gitelson, A. L. Nguy-Robertson, and Y. Peng, "Comparison of different vegetation indices for the remote assessment of green leaf area index of crops," *Remote Sensing of Environment,* vol. 115, no. 12, pp. 3468-3478, 2011/12/15/ 2011, doi: https://doi.org/10.1016/j.rse.2011.08.010.

[15] A. Mateo-Sanchis, M. Piles, J. Muñoz-Marí, J. E. Adsuara, A. Pérez-Suay, and G. Camps-Valls, "Synergistic integration of optical and microwave satellite data for crop yield estimation," *Remote Sensing of Environment,* vol. 234, p. 111460, 2019/12/01/ 2019, doi: https://doi.org/10.1016/j.rse.2019.111460.

[16] C. Lesk, P. Rowhani, and N. Ramankutty, "Influence of extreme weather disasters on global crop production," *Nature,* vol. 529, no. 7584, pp. 84-87, 2016.