MODEL DEVELOPMENT TO PREDICT CROP YIELD FOR ILLINOIS AND TEXAS STATES ECE/CS 498 DSG Final Project Spring 2020

Rishabh Gupta
rishabh7
Agricultural and Biological Engineering

Abhinavi Madireddy
am49
School of Information Management

Yasaswi Boyapati
yasaswi2
School of Information Management

***Abstract*—The researchers in cropping system modeling domain are intrigued by the accuracy of machine learning techniques compared to process-based convoluted crop model for predicting crop yield. That is why a several machine learning techniques are being explored to understand the interaction of crop-soil-climate nexus to better predict the crop yield. This research project is planned to explore several machine learning models to predict the crop yield based on weather parameters. In this project, daily time series of precipitation, maximum and minimum temperature are used to derive other significant crop effecting parameters like growing degree days, seasonal precipitation, wet days, days above and below optimum temperature for crop to predict the annual corn and soybean yield. The models are primarily developed for two states Illinois (102 counties) with data from 1950-2019 and Texax (254 counties) with data from 1968-2019 to predict the crop yield at county scale.**

I. INTRODUCTION

The drastically changing climate and ever-rising population has placed the world food security at the risk. [1]. Therefore, it has become necessary to keep tracking better methodologies to predict the agricultural production. Such predictions can be used to determine the probability that whether there would be enough food in future to feed the expected population.

Earlier researches in the crop modeling domain, used a combination of the process-based crop models and global/regional climate models (GCMs/RCMs) future climatic projections to estimate the climate change impact on agricultural production [2]–[7]. These crop models incline to be convoluted as they require a lot of data related to crop management, weather information, soil profile texture details [8]. Moreover, these models need site specific calibration for the crops which necessitate so many years of experimented field observations. The calibrating such models itself is another challenge as they have different set of parameters to calibrate the different sub-processes. This is quite extensive process which requires intensive field work and data collection at different time scale (hourly, daily, weekly, etc.) for several years. Though, all the work on field and data collection is notable since these models predicts the crop growth at a very high accuracy; however, the results from these models are more research oriented than practical applicability due to their complex structure.

Researchers these days are intrigued with the machine learning (ML) approaches due to their higher accuracy in the agricultural production prediction [9], [10]. Now-a-days, due to advancement of technology, it has become easier to collect data through a variety of sensors without harming environment which enable researchers to understand the relationship between crop, soil, and climate dynamics. In the agriculture world, the term ‘digital agriculture’ is getting popular due to emergence of data driven approaches and ML techniques [11]. Several researchers started with simple linear regression model to predict the crop behavior [8], [12], [13]. There are several studies which used complex ML techniques like random forest, multiple linear regression, support vector machines, artificial neural network (ANN), etc. to develop complicated mathematical relationship with crop-soil-climate nexus [8], [11], [14]–[22]. All these researches tried to find best model and parameters to explain the crop yield.

Based on analyzed research gap, the project was planned to demonstrate the applicability of different machine learning techniques to predict the crop yield based on weather parameters. The models developed in the project are primarily intended to predict the yield of cash crops like corn and soybean at the county scale of State Illinois and Texas.

II. DATA AND METHODS

A. Data

The historical observed precipitation and temperature (maximum and minimum) were downloaded for each county of Illinois (total of 102 counties) and Texas (total of 254 counties) states from the Midwestern Regional Climate Center online data portal (Table I). The online data portal does not provide any straightforward way where all the county-wise data for both the states for all the climate variables could be downloaded. Hence, one by one each county file was downloaded seperately for each climatic varible was downloaded which contained daily data (Precipitation/Minimum

Temperature/Maximum Temperature) recorded at various stations between 1950-2019 for Illinois and 1968-2019 for Texas State. The crop yields (corn and soybean) were downloaded from the National Agricultural Statistics Service, United States Department of Agriculture for both the states(I).

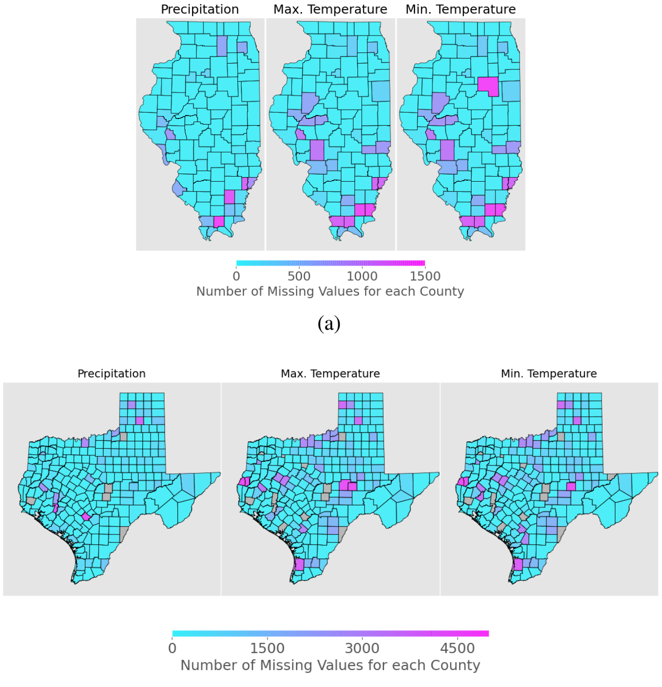
TABLE I: Sources of downloaded data.

Dataset	Sources
Precipitation (mm) Maximum Temperature (°C) Minimum Temperature (°C)	Midwestern Regional Climate Center https://mrcc.illinois.edu/
Corn Yield (kg/ha) Soybean Yield (kg/ha)	National Agricultural Statistics Service https://quickstats.nass.usda.gov/

For further processing, data cleaning is a necessary step that allows us to parse the data and check for any missing values. Since here all the data downloaded was raw; hence, the priority was to develop a single file for each climate variable from the county-wise files. A brief summary of for handling missing data can be explained in two points- firstly, since each county has several stations, the data was averaged for all the stations to generate the time series of each climate variable. Secondly, all the time series for different counties were combined into a single file for each climatic variable. Similarly, the data was parsed for Texas state. All the units of data were converted to International units from the US unit system.

Thereafter, the weather data was checked for missing values. The precipitation data were available for all counties of the states with some missing values whereas there is no data recorded of minimum and maximum temperature for few counties (Fig. 1). Since the temperature does not change much at the same latitude, the data adjacent to missing counties (same latitude, which means either left or right county) was used as the proxy observation for minimum and maximum temperature. The algorithm to handle missing data was developed in such a way that it could take the next closest county data if the closest county data was also missing. For the precipitation data, a separate algorithm was developed which replaces the missing value with the last five year average value of the particular month. In this way, the algorithm considered the seasonality of the precipitation as well. In Illinois, the corn and soybean are generally planted during April to mid-May and are harvested during mid-September and mid-October, the period from mid-April to mid-October was considered as crop growing season for the model development [23]. For Texas, the crop growing season was considered a month earlier from mid-March to mid-September [23]. Based on daily observed weather data, the parameters listed in (Table II) were calculated using daily climate data for crop growing season of respective states. All these variables are generally used to determine/explain crop behavior.

From Fig. 3, it can be seen that there were a very few counties in Texas, where the crop yield data was available; hence, the counties with more than 20 years of data available were selected for modeling purpose.



(a)

(b)

Fig. 1: Number of missing daily data for climatic variables in Illinois and Texas

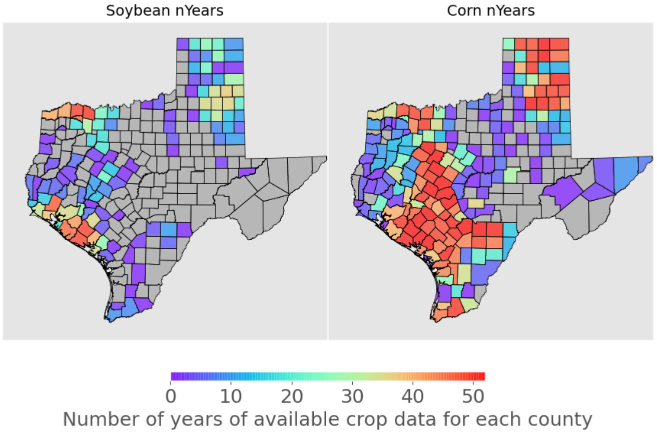


Fig. 2: Number of years of available crop data in Texas

B. Methods

1) Linear Regression: Linear regression is a linear model, a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, output variable can be calculated from a linear combination of the input variables. The method is referred to as simple linear regression when there is a single input variable and as multiple linear regression when there are multiple input variables.

The representation is a linear equation that combines a specific set of input values and the predicted output for that set of input values. The linear equation assigns one scale

TABLE II: Features/parameters calculated from daily precipitation, minimum and maximum temperature values.

Features/parameters	Description
Corn	Corn yield in kg/ha
Soybean	Soybean yield in kg/ha
Year	Illinois: 1950-2019, Texas:1968-2019
Latitude	Latitude of county centroid
Longitude	Longitude of county centroid
SeasonalPrep	Sum of daily precipitation for the season (mm)
SeasonalAvgTmax	Average daily max. temperature for the season (°C)
SeasonalAvgTmin	Average daily min. temperature for the season (°C)
SeasonalTmax	Max. temperature in the season (°C)
SeasonalTmin	Min. temperature in the season (°C)
GDD	Growing Degree Days (°C) = $\sum$ [(Tmax - Tmin)/2 - 10] for the season [If Tmax >30: Tmax = 30, If Tmin <10: Tmin = 10]
WetDays	Number of rainy days in the season
DaysA30	Number of days in the season [Tmax>30 (°C)]
DaysB10	Number of days in the season [Tmin<10 (°C)]
DaysB_2	Number of days in the season [Tmin<-2 (°C)]

factor to each input value or column, called a coefficient, and represented by the capital Greek letter Beta (β). One additional coefficient is also added, often called the intercept or the bias coefficient.

In a simple regression problem (a single x and a single y), the form of the model would be: $y = B_0 + B_1x$. In higher dimensions, the line is called a plane or a hyperplane. A coefficient value of zero removes the influence of the input variable on the model. With simple linear regression, statistics are used to estimate the coefficients. It requires calculation of statistical properties from the data such as means, standard deviations, correlations, and covariance. With multiple linear regression, Ordinary Least Squares is used to estimate the values of the coefficients. The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals.

2) Random Forest: The fundamental idea behind a random forest is to combine the predictions made by many decision trees into a single model. Individually, predictions made by decision trees may not be accurate but combined, the predictions will be closer to the true value on average. It is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique known as bagging. Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement. Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples: For b = 1, ..., B: 1. Sample, with replacement, n training examples from X, Y; call these Xb, Yb. 2. Train a classification or regression tree fb on Xb, Yb and a prediction is recorded for each sample. 3. Finally, ensemble prediction is calculated by averaging the predictions of the above trees producing the final prediction. The Random Forest requires very few pieces of feature engineering and is easy to use because the default hyperparameters often produce a good prediction result. Additionally, It significantly lowers the risk of overfitting by averaging several trees. The algorithm of Random Forest has an in-built validation mechanism named Out-of-bag (OOB) score.

3) Gradient Boosting Regression Tree: ”Boosting” is a way of merging multiple simple models into a single composite model. Since simple models are added together while keeping existing trees in the model unchanged, boosting is known as an additive model. The term ”gradient” in ”gradient boosting” comes from the fact that the algorithm uses gradient descent to minimize the loss. The loss function is generally the squared error (particularly for regression problems).

Decision tress are used as the weak learners in gradient boosting. Each internal node of the tree representation denotes an attribute and each leaf node denotes a class label. Gradient boosting Regression trains a weak model that maps features to difference between the current prediction and the known correct target value. The residual predicted by a weak model is added to the existing model input and thus this process nudges the model towards the correct target. Repeating this step improves the overall model prediction.

Few parameters that can be tuned to obtain the best output from the algorithm implementation include number of estimators (denoted as n_estimators., default value is 100), subsample (denoted as subsample, default value is 1.0), learning rate (denoted as learning_rate, default value is 0.1), criterion (denoted as criterion, default value is friedman_mse) and number of Iteration no change (denoted by n_iter_no_change, default value is None). Gradient Boosting Regression generally provides better accuracy. It requires minimal data preprocessing, that results in faster implementation of the model with lesser complexity.

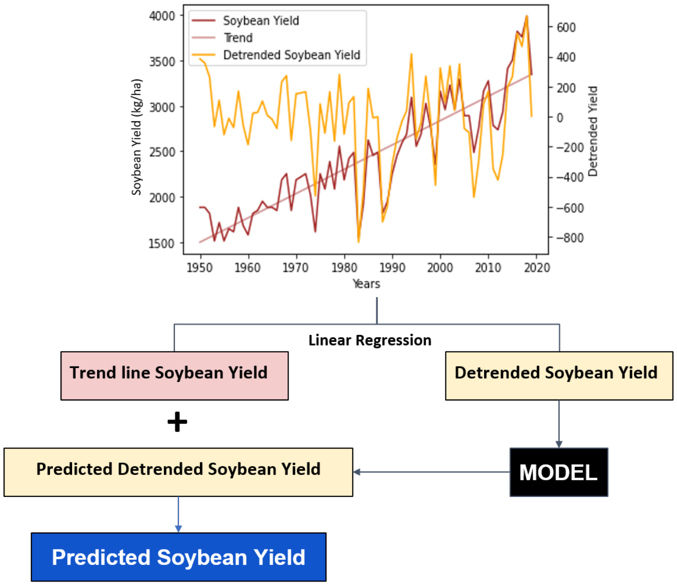


Fig. 3: Flow chart to demonstrate detrended modeling approach

4) Detrended Modeling Approach:

C. Cross validation and accuracy assessment

Cross- Validation: Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. The procedure is often called k-fold cross-validation. A specific value for k can be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. Cross-validation is primarily used to estimate the skill of a machine learning model on unseen data. That is, use a limited sample to estimate how the model is expected to perform when used to make predictions on data that is not used during the training of the model. The general procedure of Cross-validation is to shuffle the dataset randomly, split the dataset into k groups and then for each unique group take a group as a test data set and take the remaining groups as a training data set, fit a model on the training set and evaluate it on the test set, retain the evaluation score and discard the model, finally summarize the skill of the model using the sample of model evaluation scores. Cross-validation generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

Coefficient of determination: The coefficient of determination denoted by R^2 is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable. The coefficient of determination ranges from 0 to 1. An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable. An R^2 of 1 means the dependent variable can be predicted without error from the independent variable. And an R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in Y is predictable from X; an R^2 of 0.20 means that 20 percent is predictable; and so on. The coefficient of determination (R2) for a linear regression model with one independent variable is: $R^2 = (1/N) * \sum[(x_i - \bar{x}) \times (y_i - \bar{y})]/(\sigma_x * \sigma_y)^2$, where N is the number of observations used to fit the model, $\sum$ is the summation symbol, x_i is the x value for observation i, $\bar{x}$ is the mean x value, y_i is the y value for observation i, $\bar{y}$ is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

III. Results

Model	Corn Yield Accuracy	Soybean Yield Accuracy
Linear Regression	82%	78%
Random Forest	68%	64%
Gradient Boosting	95%	93%

TABLE III: Accuracies achieved with normal Illinois crop and weather data.

Model	Corn Yield Accuracy	Soybean Yield Accuracy
Linear Regression	86%	86%
Random Forest	-ve values	-ve values
Gradient Boosting	89%	87%

TABLE IV: Accuracies achieved with de-trended Illinois crop and weather data.

Model	Corn Yield Accuracy	Soybean Yield Accuracy
Linear Regression	67%	46%
Random Forest	69%	37%
Gradient Boosting	86%	50%

TABLE V: Accuracies achieved with normal Texas crop and weather data.

IV. Discussion

TODO: In this section, we recommend that you (i) draw any relevant conclusions from the dataset that can be substantiated by your results, (ii) discuss the implications of your work in real-world context, and (iii) mention limitations/next steps with your project.

From the tables in results section, it is evident that in almost all models, we see accuracy of corn models came out to be greater than soybean models. This relatively higher accuracy might be of reason that many of the variables are highly correlated to corn yield value than soybean yield value like SeasonalTmax, DaysA30, Year. It can also be seen that Detrended data performed better in Linear Regression but went low in Gradient Boosting. Which could be due to re-trending of data resulted from linear regression after applying gradient boosting method on detrended yield values.

In the case of Texas, due to limited amount of data available, all the models were giving out lower accuracies. For the same reason, we decided not to go ahead with detrended data here.

Based on the results and context, although Gradient Boosting Regression predicted the yield values with better accuracy, we recommend using Linear regression. Because Linear Regression also achieved relatively impressive accuracy with detrended data and is simple to work with and understand when compared to Gradient Boosting.

V. Member Contributions

TODO: In this section, briefly comment on the individual group member contributions to the final project.

VI. Acknowledgment

(optional). Feel free to acknowledge anyone who helped you in designing and implementing your project.

References

[1] F. (Food and A. Organization), “Global agriculture towards 2050,” in *In High Level Expert Forum - How Feed World*, vol. 2050, 2009, pp. 1–4.

[2] M. Moriondo, C. Giannakopoulos, and M. Bindi, “Climate change impact assessment: the role of climate extremes in crop yield simulation,” *Climatic change*, vol. 104, no. 3-4, pp. 679–701, 2011.

[3] J. E. Olesen *et al.*, “Impacts and adaptation of european crop production systems to climate change,” *European Journal of Agronomy*, vol. 34, no. 2, pp. 96–112, 2011.

[4] J. Knox *et al.*, "Climate change impacts on crop productivity in africa and south asia," *Environmental Research Letters*, vol. 7, no. 3, p. 034032, 2012.

[5] S. Asseng *et al.*, "Uncertainty in simulating wheat yields under climate change," *Nature climate change*, vol. 3, no. 9, pp. 827–832, 2013.

[6] L. Ye *et al.*, "Climate change impact on china food security in 2050," *Agronomy for Sustainable Development*, vol. 33, no. 2, pp. 363–374, 2013.

[7] C. Rosenzweig *et al.*, "Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison," *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3268–3273, 2014.

[8] M. Kaul, R. L. Hill, and C. Walthall, "Artificial neural networks for corn and soybean yield prediction," *Agricultural Systems*, vol. 85, no. 1, pp. 1–18, 2005.

[9] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Computers and electronics in agriculture*, vol. 151, pp. 61–69, 2018.

[10] K. G. Liakos *et al.*, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, p. 2674, 2018.

[11] N. Gandhi *et al.*, "Rice crop yield prediction in india using support vector machines," in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, 2016, pp. 1–5.

[12] R. P. Singh *et al.*, "Crop yield prediction using piecewise linear regression with a break point and weather and agricultural parameters," Apr. 20 2010, uS Patent 7,702,597.

[13] V. Sellam and E. Poovammal, "Prediction of crop yield using regression analysis," *Indian Journal of Science and Technology*, vol. 9, no. 38, p. 5, 2016.

[14] S. Fukuda *et al.*, "Random forests modelling for the estimation of mango (mangifera indica l. cv. chok anan) fruit yields under different irrigation regimes," *Agricultural water management*, vol. 116, pp. 142–150, 2013.

[15] A. González Sánchez *et al.*, "Predictive ability of machine learning methods for massive crop yield prediction," 2014.

[16] S. S. Dahikar and S. V. Rode, "Agricultural crop yield prediction using artificial neural network approach," *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, vol. 2, no. 1, pp. 683–686, 2014.

[17] M. Paul, S. K. Vishwakarma, and A. Verma, "Analysis of soil behaviour and prediction of crop yield using data mining approach," in *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2015, pp. 766–771.

[18] D. Gouache *et al.*, "Agrometeorological analysis and prediction of wheat yield at the departmental level in france," *Agricultural and Forest Meteorology*, vol. 209, pp. 1–10, 2015.

[19] X. E. Pantazi *et al.*, "Wheat yield prediction using machine learning and advanced sensing techniques," *Computers and Electronics in Agriculture*, vol. 121, pp. 57–65, 2016.

[20] Y. Everingham *et al.*, "Accurate prediction of sugarcane yield using a random forest algorithm," *Agronomy for sustainable development*, vol. 36, no. 2, p. 27, 2016.

[21] J. H. Jeong *et al.*, "Random forests for global and regional crop yield predictions," *PLoS One*, vol. 11, no. 6, 2016.

[22] A. Crane-Droesch, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture," *Environmental Research Letters*, vol. 13, no. 11, p. 114003, 2018.

[23] U. S. D. of Agriculture. Statistical Reporting Service, U. S. Science, and E. Administration, *Usual planting and harvesting dates for US field crops*. US Department of Agriculture, Statistical Reporting Service, 1949, no. 628.