Crop Yield Prediction using Supervised Statistical and Machine Learning Techniques

Prajna Kandarpa
Mechanical and Mechatronics Engineering
University of Waterloo
Waterloo, Canada
Email: spspkand@uwaterloo.ca

***Abstract*—Neural Networks and Multiple Adaptive Regressive Splines (MARS), were used to predict crop yield for cereals for the country of India. Historical annual cereal yield data from 1960-2010 for the country of India was used as target data with seasonal mean temperatures, precipitation and arable land size as predictors. The two models were trained with this data and their predictions were compared with the actual yields as well as the performance of more complex crop models.**

I. Introduction

Climate change has brought about increasingly chaotic weather patterns all over the planet making national economic and resource planning difficult for researchers, independent contractors and government organizations. These entities are therefore paying increased attention to the crop forecasts of major food-exporting countries as well as to their own domestic food production. Given the increased volatility of food markets and the rising incidence of climatic extremes affecting food production, food price spikes may increase in prevalence in future years [1]. As such, complex models that can simulate ensemble effects from climate and soil conditions while accounting for geographical variability have assumed paramount importance. Crop growth models perform an abstraction of the dynamic mechanistic of the plants physiological stages by fitting them into a mathematical model [2].

Recent research in this area has sought to build statistical models that perform just as well at forecasting yields as the complex models. Rauff (2015) [3] talk about how crop models can be used to understand the effects of climate change such as elevated nitrogen levels, CO2 levels, temperature and precipitation changes on crop development, growth and yield. For example, sudden onset of warm and humid conditions can lead to plant disease proliferation and cause crop yield loss and financial setbacks for farmers and geographical regions. It is a difficult task to produce a comprehensible, operational representation of a part of reality, which grasps the essential elements and mechanisms of that real world system and even more demanding, when the complex systems encountered in environmental management [3].

A few kinds of crop models have been developed so far, ranging from empirical models to explanatory models. A widely used approach to this prediction problem is to rely on numerical models that emulate the main processes of crop growth and development. Lobell et al (2010)[4] state that these models are typically developed and tested using experimental trials and thus offer the distinct advantage of leveraging decades of research on crop physiology and reproduction, agronomy, and soil science, among other disciplines. Yet these models also require extensive input data on cultivar (plant varieties produced by selective breeding), management, and soil conditions that are unavailable in many parts of the world.

The presence of all the data required for these models still requires calibration, which is difficult due to the high amount of uncertainty in the model parameters. Often, parameter uncertainty is ignored in such complex models and parameter values are picked through subjective deliberation. Iizumi et al. (2009) estimated distributions of parameter values from Markov Chain Monte Carlo techniques with the widths of the distributions reflecting the inability of historical rice, maize yield datasets to effectively constrain parameter values. They found that parameter uncertainties translated to larger uncertainties in projecting yield responses to climate change. They concluded that uncertainty of climate change impact assessment on crop yield may increase if future climate projections are fed to crop models with parameters optimized under current climate conditions [5].

Statistical models, on the other hand, trained with historical crop yield data, and made of relatively simple linear regression ensembles provide a common alternative to the process based models described above. These methods include the following:

1) time series methods - based on time series data from a single point or geographical area.
2) panel methods - these methods rely on variations in both time and feature space.
3) cross sectional methods - based solely on variations in space.

Time series methods are often limited by availability of data but have the advantage of capturing the variability for the chosen location. Panel and cross-sectional methods assume common parameter values across all locations and are prone to errors from omission of features like soil conditions or fertilizer uptake. The main advantages of statistical models are their limited reliance on field calibration data, and their transparent assessment of model uncertainties. For example, if a model does a poor job of representing crop yield responses to climate, this will be reflected in a low coefficient of

determination (R^2) between modeled and observed quantities, as well as a large confidence interval around model coefficients and predictions. Although process-based models could in theory be accompanied with similar statistics, in practice they rarely are [4]. A few shortcomings with statistical models include co-linearity problems and assumptions of stationarity of the underlying ecological processes, which is an extremely important problem due to rapidly changing ecologies on the planet as a result of global warming.

Lobell et al.[4] recognized the need for a systematic evaluation of statistical methods in predicting crop yields to climate and recommended that efforts be made to determine the specific conditions under which their predictions are likely to be misleading. This would help in quantifying the common errors that arise from using these simpler if imperfect approaches. Researchers [6] [7] have found that statistical models work best at large spatio-temporal scales like national or regional levels. Other researchers have highlighted the difference between understanding the impact of climate change on crop growth and on crop yields. Predicting crop yields involves understanding predictors that aren't reflected in the physiological crop growth models explaining their inability to reliably predict crop yields. This problem maybe considered as one of dimensionality reduction where multiple predictors need to be aggregated at regional, provincial and national levels and temporal scales like monthly or seasonal climatic variable changes.

A. Objective

The main goal of this report is to evaluate the accuracy of context unaware statistical models at accounting for the uncertainty found in the task of crop yield prediction. Statistical models that use Neural Networks and Multiple Adaptive Regressive Splines are to be built using historical annual crop yield, seasonal temperature and precipitation levels and assess their accuracy at predicting yields using measures such as error bounds, confidence intervals on predictions and compare these measures of variability with similar measures obtained by crop growth models. Many researchers [8] [1] [2] have found that statistical models can explain variability in crop yields due to climate change better than crop growth models do. The important features that explain this variability were found to be seasonal variations in weather, precipitation, fertilizer intake and crop breeding effects. This study aims to make similar inferences about the importance of climatic and geophysical factors on crop yields based on the results obtained from the Neural Network and MARS models.

II. Data and Methods

A. Datasets

Data was obtained from the World Bank's data on socioeconomic indicators for all countries in the world [9]. This data includes historical agricultural data like crop yields, fertilizers used, agriculture related carbon dioxide, nitrogen dioxide emissions and extensive climate data going back to 1901. The data was accessed using their web interface and via the R packages *rWBclimate* [10] and *WDI* [11].

The crop yield data obtained was for annual cereal production in India from 1961 to 2013. Cereals are a plant family consisting primarily of wheat, oats and corn. Monthly temperature and precipitation data at a national scale was also obtained from 1961-2013 and a subset of this data corresponding to primary agricultural seasons in India was taken to be used for training purposes. A summary of the collected data is available in Table I

Statistic	N	Mean	St. Dev.	Min	Max
land	53	100,044,223.600	3,429,131.695	92,239,016	106,613,208
year	54	1,986.500	15.732	1,960	2,013
yield	53	1,759.542	633.236	854.400	3,010.000
machines	49	16.291	1.577	12.802	19.622
fertilizer	12	144.211	27.629	100.329	180.748
temp_rabi	54	19.824	0.536	18.665	21.337
temp_kharif	54	25.973	0.294	25.481	26.689
rain_rabi	54	12.364	2.062	8.606	17.296
rain_kharif	54	186.631	18.292	140.926	218.241

TABLE I
SUMMARY OF DATA COLLECTED FOR THIS STUDY

B. Methods

The use of complex classification and regression models is becoming more and more commonplace in science, finance and a myriad of other domains. Regression is concerned with modeling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model. Regression methods are a workhorse of statistics and have been cooped into statistical machine learning. This may be confusing because one can use regression to refer to the class of problem and the class of algorithm. Really, regression is a process.

Mathematically, we can define the process of Regression modeling as follows. Suppose, we have an output Y and a series of inputs or predictors (usually assumed to be independent variables).

$$X_1, X_2, ...X_n$$

Then, the goals of a regression model are multiple:

1) examine the relationship between inputs and outputs – Do they tend to vary together? What does the structure of the relationship look like? Which inputs are important?
2) Given a new set of predictor values $X_1^*, ...X_p^*$, what can be said about an unseen Y^*?
3) Regression tools often serve as a building block for more advanced methodologies - Smoothing by local polynomials, for example, involves fitting lots of regression models "locally", while iteratively fitting weighted regressions is at the heart of the standard computations for generalized linear models

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such

algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable; for example, correlation does not imply causation. In a narrower sense, regression may refer specifically to the estimation of continuous response variables, as opposed to the discrete response variables used in classification. The case of a continuous output variable may be more specifically referred to as metric regression to distinguish it from related problems.

The dataset summarized in Table I was placed into a data frame in R. A data frame is a glorified matrix that allows easy indexing for splitting and identifying data pertaining to separate trials or conditions. A sample of the dataset is provided in Table II. For the purposes of this problem, the *yield* column from Table II is the desired target variable we are trying to predict. All the other variables are to be sued as predictors.

A convenience function provided by *caret*, *nearZeroVar* was then used to determine a few non useful predictor variables and these were excluded from the predictors for testing and training datasets. These so-called near zero-variance predictors can cause problems during resampling for some models such as linear regression [12]. Other variables like agricultural machines and fertilizer intakes were dropped from the final dataset since sufficient data was not enough to account for the entire temporal range being considered for this study: 1961-2013. Although, methods that either help extrapolate missing data in sequences like Markov Chain Models and statistical models that can work with incomplete data exist, their applicability was deemed to be out of scope for this study in light of time constraints.

	land	year	yield	machines	fertilizer	temp_rabi	temp_kharif	rain_rabi	rain_kharif
1	99,190,000	2,013	2,963.400		157.522	18.665	25.560	11.828	197.899
2	97,440,000	2,012	3,010		164.783	18.906	25.742	14.045	192.931
3	100,625,700	2,011	2,860.700	18.535	180.748	19.837	25.938	13.910	215.299
4	100,075,800	2,010	2,676.400	19.052	179.036	21.337	26.634	11.362	173.963
5	97,171,600	2,009	2,580.800	19.622	167.457	20.006	26.197	13.955	182.970
6	101,155,500	2,008	2,637.900	17.906	153.349	20.187	26.386	13.174	197.289

TABLE II
SAMPLE ROWS FROM THE DATA FRAME

Table II, curiously shows two sets of features for rainfall and weather respectively: rain_kharif, rain_rabi, temp_kharif, temp_rabi. The agricultural crop year in India is from July to June. The Indian cropping season is classified into two main seasons-(i) Kharif and (ii) Rabi based on the monsoon. The kharif cropping season is from July October during the south-west monsoon and the Rabi cropping season is from October-March (winter). The crops grown between March and June are summer crops. Pakistan and Bangladesh are two other countries that are using the term kharif and rabi to describe about their cropping patterns. The terms kharif and rabi originate from Arabic language where Kharif means autumn and Rabi means spring [13]. Thus, the 4 climatic variables in the datasets correspond to the cultivation seasons are are means of the monthly values over the respective months in which they are applicable.

C. Data Pre-processing

In the field of statistical data analysis, one of the first tasks is to determine how much of the finite dataset is to be used for model training while ensuring a certain portion of the dataset is kept aside for testing the efficacy of the model after training. Thus, it is important to ensure that a model being trained never gets exposed to the split testing /validation dataset. This is a very good measure of determining how well the model would perform when being used in real life. The main function of the test split dataset is to compare and evaluate performance across models, as most statistical models are actually combinations of localized models.

The *caret* package, specifically has a *createDataPartition* function, that analyses a dataset's characteristics such as multivariate correlation and determines the most randomized way to split the dataset. For regression, the function determines the quartiles of the data set and samples within those groups. Thus, 70% of the dataset was randomly sampled to be used as the training data. The rest of the 30% of the dataset was then used to evaluate the performance of the models, hereafter known as the testing data or validation data.

D. Tuning and building models

```
Define sets of model parameter values to evaluate
for each parameter set do
    for each resampling iteration do
        Hold–out specific samples
        [Optional] Pre–process the data
        Fit the model on the remainder
        Predict the hold–out samples
    end
    Calculate the average performance across hold–out predictions
end
Determine the optimal parameter set
Fit the final model to all the training data using the optimal parameter set
```

Fig. 1. Standard operating procedure to tune a model's parameters [14]

The general process that needs to be followed to do efficient parameter tuning while building a model is shown in Figure. 1. The *caret* package has several functions that streamline the process of model building, tuning and evaluation. The *train* function can be used to

- Evaluate, using resampling the effects of model parameter tuning on performance
- Choose the optimal model across these parameters
- Estimate model performance from a training set[14]

1) Neural Networks: Artificial Neural networks are computing systems made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output. FOr this study, one hidden layer was used with the number of hidden layer neurons being a parameter that is tuned for. An ANN maybe described by two parameters, size and decay. Its performance can be evaluated using regression plots, and RMSE and the coefficient of determination, R^2, which is a measure of how well the model was able to predict each sample.

2) Multiple Adaptive Regressive Splines (MARS): MARS is a form of regression analysis introduced by Jerome H. Friedman in 1991. It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models nonlinearities and interactions between variables. The MARS model is a weighted sum of Basis functions,

$$f(x) = \sum_{1=1}^{k} c_i B_i(x)$$

[15]. c_i is a constant coefficient. Each basis function can take three different forms - a constant, a hinge function and a product of two or more hinge functions.

MARS models may also be evaluated using the standard regression measures RMSE and the coefficient of determination, R^2

III. Results and Discussion

Figures 2 and 4 show the parameter tuning performance and training performance of the neural network used on the dataset. The NN model had its best performance with 5 hidden layer neurons and a weight decay of 0.1. The MARS model

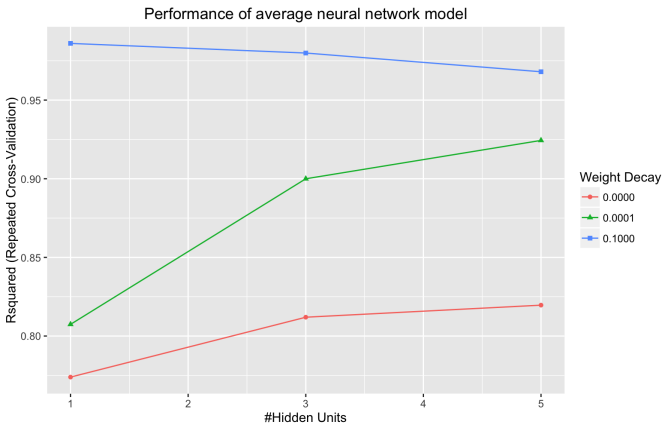

Fig. 2. Parameter tuning results for Neural network

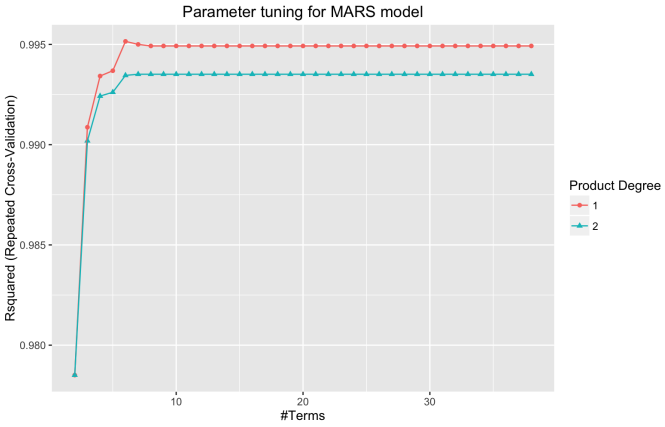

Fig. 3. Parameter tuning results for MARS

As can be seen in Figures 4 and 6, both models fit the training data well with RMSE errors staying pretty low and the coefficient of regression fit, R^2, at around 0.98. One might be inclined to argue that such good of a fit might point towards over-fitting, which would very well have been true if not for the k-fold cross validation used during the training process which ensures none of the model parameters have been biased for a specific data point.

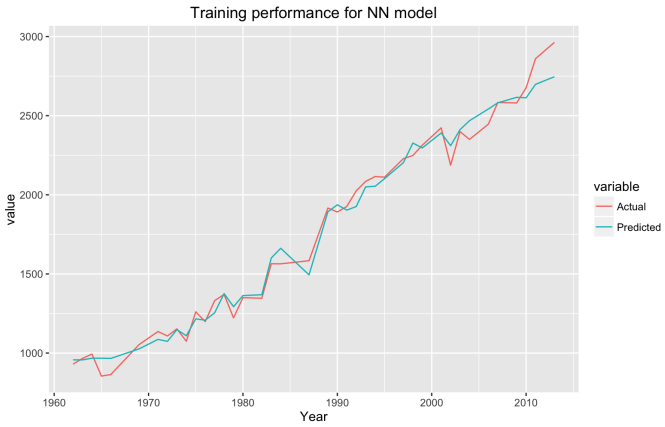

Fig. 4. Training results for NN

Figures 5 and 7 show the prediction accuracy of the models on the validation dataset. Once again, the predictions closely followed the actual values proving the efficacy of these models in predicting crop yield. The RMSE and R^2 results for the training, Table III, and testing, Table IV, data-splits for both MARS and Neural Networks have been provided. As is obvious from them, both models had extremely similar performance in fitting the validation data to the model. The low RMSE values, which range from 4-8% of the mean yield value, indicate really good performance for these models.

IV. Summary and Conclusions

In conclusion, we found that the models trained here, based on neural networks and Multiple Adaptive Regressive Splines,

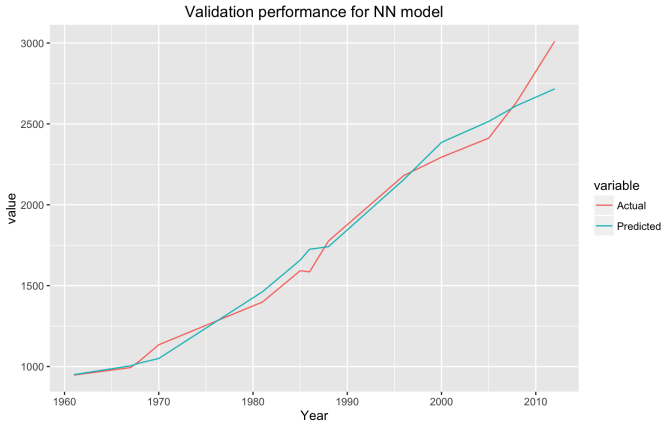

Fig. 5. Validation results for NN

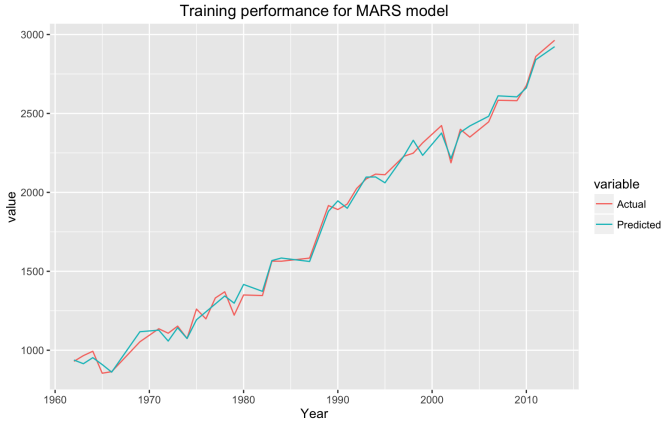

Fig. 6. Training results for MARS

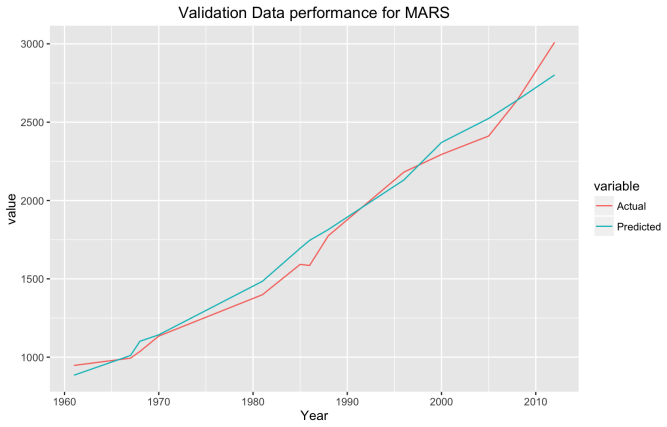

Fig. 7. Validation results for MARS

res	rmse	r2
AvgNN	71.8543768205759	0.986877633023676
MARS	41.7083829642748	0.995445879757786

TABLE III
RMSE AND R^2 RESULTS FOR TRAINING

res	rmse	r2
AvgNN	105.0088152187575	0.975009493599273
MARS	95.1923222679583	0.981305690867502

TABLE IV
RMSE AND R^2 RESULTS FOR TESTING

were able to accurately represent the behavior of the annual crop yields for cereals in India. An assessment of the variables deemed to be important by these models is provided in Figures 8 and 9.

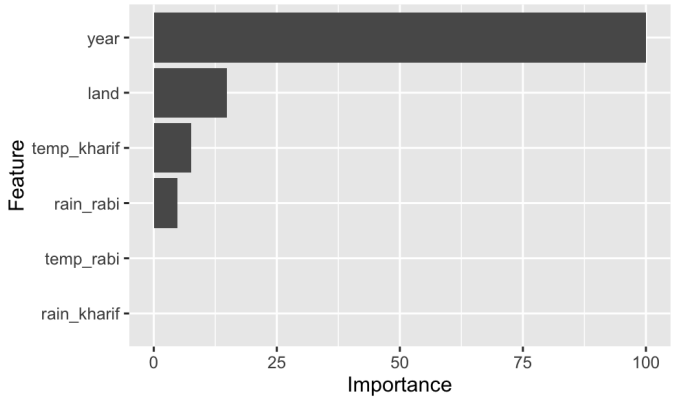

Fig. 8. Important variables in MARS model

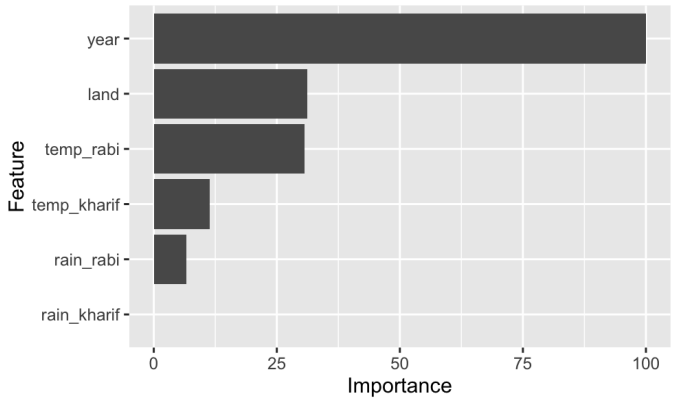

Fig. 9. Important variables in NN model

These results indicate that the year variable was most important in predicting future crop yields. This makes sense when one can think about the year acting as a proxy for the actual factor driving crop yield higher every year - population. It would be very interesting to run the same analysis using country population instead of the year and verify this claim. Other noticeable aspects include the negligible importance of the rain variables in predicting crop yields. This maybe attributed to large regional variations in precipitation which are lost when one considers aggregated weather data at national scales. It is very likely that the regions responsible for most of the cereal production received much higher rainfall than represented by the seasonal mean values used in this study.

Previous works have suggested that climatic variables can account for most of the variations in crop yields. Lobell et al.(2010) found that the performance of statistical models differed by climate variable and spatial scale, with time-series statistical models ably reproducing site-specific yield response to precipitation change, but performing less well for temperature responses. The models based on multiple sites were also much less sensitive to the length of historical period used for training. For all three statistical approaches, the performance improved when individual sites were first aggregated to country-level averages. Results suggest that statistical models, as compared to CERES-Maize, represent a useful if imperfect tool for projecting future yield responses, with their usefulness higher at broader spatial scales. It is also at these broader scales that climate projections are most available and reliable, and therefore statistical models are likely to continue to play an important role in anticipating future impacts of climate change [4].

Thus, it is recommended that better national scale predictor aggregates be obtained primarily from the major regions responsible for cereal production in India, including temperature, rainfall and fertilizer intake to obtain much better results as suggested by Lobell et. al. Other recommendations include adding energy usage by the agricultural sector and irrigation levels, using techniques that allow extrapolation of missing data for the variables that were dropped during the data preprocessing stage of this study like agricultural machines and fertilizer uptake. Researchers have found fertilizer uptake, to be an especially useful factor for predicting crop yields as it is also linked to popular crop strains and breeding practices that may vary regionally. Gonzalez-Sanchez et. al found that M5-prime regression trees performed really well, which can be another model to build using this dataset [2].

Finally, it is necessary to point out that this work deals only with comparing the predictive accuracy of the above-mentioned techniques. Several factors like model structure, knowledge representation, implementation cost and training time affect the performance of machine learning techniques and further research needs to be done to compare the characteristics of the ML models with agricultural planning factors.

ACKNOWLEDGMENT

The author would like to thank H.R. Tizhoosh, Professor, Systems Design Engineering, University of Waterloo for his suggestions about ensuring enough data is available to perform the analysis conducted here but on a regional or smaller spatial scale. The original intent of this report was to use regional climate, soil and fertilizer data and perform analysis at a much smaller spatial scale with the same temporal scale. However, technical difficulties prevented such analyses and the spatial scale had to be increased to the national level to proceed.

REFERENCES

[1] T. Iizumi, H. Sakuma, M. Yokozawa, J.-J. Luo, A. J. Challinor, M. E. Brown, G. Sakurai, and T. Yamagata, "Prediction of seasonal climate-induced variations in global food production," *Nature climate change*, vol. 3, no. 10, pp. 904–908, 2013.

[2] a. Gonzalez-Sanchez, "Predictive ability of machine learning methods for massive crop yield prediction," *Spanish Journal of Agricultural Research*, vol. 12, no. 2, pp. 313–328, 2014. [Online]. Available: http://revistas.inia.es/index.php/sjar/article/view/4439

[3] K. O. Rauff and R. Bello, "A Review of Crop Growth Simulation Models as Tools for Agricultural Meteorology," no. September, pp. 1098–1105, 2015.

[4] D. B. Lobell and M. B. Burke, "On the use of statistical models to predict crop yield responses to climate change," *Agricultural and Forest Meteorology*, vol. 150, no. 11, pp. 1443–1452, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.agrformet.2010.07.008

[5] T. Iizumi, M. Yokozawa, and M. Nishimori, "Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: Application of a Bayesian approach," *Agricultural and Forest Meteorology*, vol. 149, no. 2, pp. 333–348, 2009.

[6] "A review on statistical models for identifying climate contributions to crop yields," *Journal of Geographical Sciences*, vol. 23, no. 3, pp. 567–576, 2013.

[7] D. B. Lobell and M. Burke, *Climate change and food security: adapting agriculture to a warmer world*. Springer Science & Business Media, 2009, vol. 37.

[8] S. Wolfram and B. L. David, "Robust negative impacts of climate change on African agriculture," *Environmental Research Letters*, vol. 5, no. 1, p. 14010, 2010. [Online]. Available: http://stacks.iop.org/1748-9326/5/i=1/a=014010

[9] W. Bank, "World development indicators — data," http://data.worldbank.org/data-catalog/world-development-indicators, (Accessed on 04/23/2016).

[10] E. Hart, *rWBclimate: A package for accessing World Bank climate data*, 2014, r package version 0.1.4.99. [Online]. Available: http://www.github.com/ropensci/rwbclimate

[11] V. Arel-Bundock, *WDI: World Development Indicators (World Bank)*, 2013, r package version 2.4. [Online]. Available: https://CRAN.R-project.org/package=WDI

[12] M. Kuhn, "Building Predictive Models in R Using the caret Package," *Journal Of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008. [Online]. Available: http://www.jstatsoft.org/v28/i05/

[13] "Cropping seasons of india- kharif rabi - arthapedia," (Accessed on 04/23/2016). [Online]. Available: http://goo.gl/lk9XQ1

[14] M. Kuhn, "Model training and tuning," http://topepo.github.io/caret/training.html, (Visited on 01/11/2016).

[15] J. H. Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, pp. 1–67, 1991.