Predicting wheat yield traits using machine learning algorithms and hyperspectral image data

Alex Whitehurst
University of Florida
Ajam1097@yahoo.com
UFID: 6469-1924

Paper Prepared for Honors Thesis
Faculty Advisor: Mostafa Reisi Gahrooei, PhD
Spring 2020

Department of Industrial & Systems Engineering
University of Florida
ABSTRACT

Predicting yield in crops has been shown to be valuable for crop management in many ways, motivating the formulation of predictive models. When data collected is often time- and labor-intensive, less costly alternatives are preferable. This study explores the usefulness of machine learning algorithms with input data obtained through UAVs equipped with hyperspectral cameras to predict yield, grain number per meter squared and cellular membrane thermostability. Dimension reduction methods such as principal components analysis (PCA) are shown to be valuable for inputs into linear and nonlinear regression methods for prediction with hyperspectral data.
# TABLE OF CONTENTS

1 INTRODUCTION .................................................................................................................. 4

2 DATA .................................................................................................................................. 4

2.1 DESCRIPTION .................................................................................................................. 4
2.2 RESPONSE VARIABLES .................................................................................................. 5
2.3 HYPERSONTAL IMAGES .................................................................................................... 6
2.4 GENOTYPE ....................................................................................................................... 7

3 METHODOLOGY .................................................................................................................. 8

3.1 MODEL TESTING & SELECTION ....................................................................................... 8
3.2 MODEL PERFORMANCE METRICS ............................................................................... 9
3.3 TESTING AND VALIDATION ........................................................................................ 10

4 RESULTS ............................................................................................................................ 10

4.1 YIELD PREDICTION ....................................................................................................... 10
4.2 GNM2 PREDICTION ...................................................................................................... 12
4.3 CMT PREDICTION .......................................................................................................... 13

5 DISCUSSION & FUTURE WORK ........................................................................................ 15

5.1 REPEATABILITY ............................................................................................................ 15
5.2 WEATHER CONSIDERATIONS .................................................................................... 15
5.3 HYPERSONTAL DATA OVER TIME ........................................................................... 16

6 CONCLUSION ...................................................................................................................... 16

7 APPENDIX .......................................................................................................................... 17

7.1 RIDGE REGRESSION ...................................................................................................... 17
7.2 PRINCIPAL COMPONENTS ANALYSIS (PCA) ................................................................. 17
7.3 PARTIAL LEAST SQUARES REGRESSION (PLS) ............................................................ 17
7.4 TREES REGRESSION ..................................................................................................... 18

8 REFERENCES ....................................................................................................................... 19
1 INTRODUCTION

The ability to predict crop yield has been shown to be valuable for crop management practices. Accurate yield prediction enables the improvement of these practices, including nutrient adjustment during the season, marketing, and preparation for harvest [1, 2]. The value of estimating crop yield and other useful crop attributes motivates the need to be able to make these evaluations at low cost. Prediction methods may involve time- and labor-intensive measurements with repeated observations throughout the season. The use of canopy reflectance and vegetation indices are included in many methods that have been used to predict wheat yield [1, 2]. A different method, namely the use of hyperspectral image data, has been shown in the prediction of maize yield to outperform vegetation indices [3], motivating the use of this data for the prediction of wheat yield. Estimation methods that are repeatable and involve rapid data collection would be valuable, particularly when data is unavailable or difficult to obtain.

With this emphasis on wheat yield, it must be noted that stress due to high temperatures is a factor that can significantly influence the growth of grain [4]. Cellular membrane thermostability has been studied and shown to be a good measurement for evaluating heat tolerance [5]. In Florida, because of high temperatures throughout the year, it is valuable to understand how crops will tolerate heat, particularly with a wheat such as the one in this study.

Given the usefulness of predicting yield and heat tolerance, this study aims to explore the ability of various machine learning algorithms with hyperspectral image data to predict three variables: yield, grain number per meter squared and cellular membrane thermostability.

2 DATA

2.1 Description

The dataset studied in this paper provides 143 records of plots from planting to harvest including various yield outputs, agricultural indicators, and hyperspectral image data taken on multiple occasions using UAV remote sensing. Pictured below is an example of the layout of plots studied in this analysis (Figure 2-1). Across all plots there are 40 different genotypes studied. This section details initial analysis of the data used as response and input variables.
2.2 Response Variables

The three response variables have been analyzed in this report with their respective predictive models. These include the yield (kg/ha), grain number per meter squared (Gn/m²) and Cellular Membrane Thermostability (CMT). Figure 2-2 shows the distribution for these three variables.

To test if these response variables were normally distributed, the ordered values were plotted against the theoretical quantiles of a normal distribution. These plots are known as qq plots. From this analysis illustrated in Figure 2-3, there is evidence that all these responses follow normal distribution. Nevertheless, some strong outliers appear on the low end of the distribution that can slightly affect the results of the models.
2.3 Hyperspectral Images

With the goal in mind of creating predictive models based on a rapid, low-cost method of collecting data on plots, the hyperspectral image data was used as the primary input variable for these models. For each plot, there are 150 frequencies, which is analogous to the 3 RGB values of a typical pixel. The data collected was averaged together to create one average value for every frequency and plot.

For a subset of the plots, we have data on four different data collection flights. Figure 2-4 shows the average curve across all plots. Each frequency is given on the x-axis with the average value of all plots given on the y-axis. To account for differences in weather or light conditions across different flights, hyperspectral data was normalized by dividing by the value of the first frequency. This normalization leads to the plot in Figure 2-5 showing a generalized decreasing trend for the values of higher frequencies as the crop gets closer to harvest date. For this paper, models were tested using one set of flights rather than a combination of all flight data because of incomplete data, correlation between flights, and complexity constraints.
2.4 Genotype

A factor suspected to potentially influence these response traits is the varying genotypes between plots. Initial analysis showed evidence that there are not significant differences between genotypes in terms of yield and other response traits. Clusters developed based on similar genotypes showed little evidence of different yields as it is shown in Figure 2-6.

To confirm this, ANOVA was performed to determine whether or not genotype is a significant factor for any of the three response variables. The report of the ANOVA test can be seen in Table 2-1. The null hypothesis that genotype is a significant factor is rejected at the 5% significance level for all three response traits. Because of this conclusion, it is assumed that genotype is not a significant factor. Genotype is thus omitted from most models and this evidence is considered in validation and testing of the models.

![Figure 2-6: Box and Whisker Plot of Yield by Genotype Cluster](image)

Table 2-1: ANOVA Table for Genotype Factor by Response Trait

<table>
<thead>
<tr>
<th>Response Trait</th>
<th>Factor</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>F-Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield</td>
<td>Genotype</td>
<td>4.0604e7</td>
<td>39</td>
<td>1.198513</td>
<td>0.233502</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>8.9474e7</td>
<td>103</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>Gn/m²</td>
<td>Genotype</td>
<td>6.4649e7</td>
<td>39</td>
<td>1.320537</td>
<td>0.135149</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>1.2930e7</td>
<td>103</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>CMT AVG</td>
<td>Genotype</td>
<td>864.9326</td>
<td>39</td>
<td>0.68008</td>
<td>0.913171</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>3358.883</td>
<td>103</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>
3 METHODOLOGY

3.1 Model Testing & Selection

Upon initial data exploration and modeling, basic linear regression (OLS) was shown to be a very poor predictor with significant overfitting due to the high dimensionality of the input data. Because of this, machine learning algorithms that involve dimension-reduction were implemented to determine the best predictive model. To test these models, k-fold cross validation methods were implemented with two model performance metrics detailed in the coming section. Because of the large number of studies using vegetation indices and canopy reflectance to predict yield, all predictive models were compared to a control model which used OLS with NDVI.

Many models were tested throughout this process, all listed in Table 3-1. Shortly after attempting linear regression, ridge regression was used to account for the high levels of correlation in the input data.

Next, dimension-reduction methods such as principal components analysis (PCA) and partial least squares (PLS) regression were tested. PCA is an unsupervised learning method that transforms large dimensions of data into “principal components” where the observations vary the most [6]. The first two principal components were chosen to include in the models as they explained 99.16% of the variance in the data (Figure 3-1). The use of PCA reduces the complexity of the problem significantly by reducing the number of predictors from 150 to just two. PLS, similarly to PCA reduces the dimensions but attempts to do so in a supervised manner [6]. In PLS, the optimal number of components was found to be one or two for the three response variables, reducing the problem complexity significantly similar to PCA.

Lastly, tree-based methods were implemented in conjunction with PCA to address non-linearity and clustering of the data. Tree regression determines splitting rules to segment the input data into regions that are used to make the prediction [6].
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Model Description</th>
<th>Inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>Linear - NDVI (Control)</td>
<td>NDVI</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>Clusters</td>
<td>Cluster, ( \forall i \in (1,9) )</td>
</tr>
<tr>
<td>Ridge Regression</td>
<td>Ridge</td>
<td>Frequency, ( \forall i \in (1,150) )</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>PCA - Linear</td>
<td>PCA1, PCA2</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>PCA - Interaction</td>
<td>PCA1, PCA2, PCA1*PCA2</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>PCA - 2nd Order Polynomial</td>
<td>PCA1, PCA2, PCA1^2, PCA2^2</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>PCA - 2nd Order Polynomial &amp; Interaction</td>
<td>PCA1, PCA2, PCA1*PCA2, PCA1^2, PCA2^2</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>PCA - 3rd Order Polynomial</td>
<td>PCA1, PCA2, PCA1^3, PCA2^3</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>PCA - 3rd Order Polynomial &amp; Interaction</td>
<td>PCA1, PCA2, PCA1*PCA2, PCA1^3, PCA2^3</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>PCA - 2nd Order Polynomial &amp; Clusters</td>
<td>PCA1, PCA2, PCA1^2, PCA2^2, Cluster: ( \forall i \in (1,9) )</td>
</tr>
<tr>
<td>Partial Least Squares</td>
<td>PLS - 1 Component</td>
<td>Frequency, ( \forall i \in (1,150) )</td>
</tr>
<tr>
<td>Partial Least Squares</td>
<td>PLS - 2 Components</td>
<td>Frequency, ( \forall i \in (1,150) )</td>
</tr>
<tr>
<td>Trees Regression</td>
<td>Trees - PCA</td>
<td>PCA1, PCA2</td>
</tr>
<tr>
<td>Trees Regression</td>
<td>Trees - Hyperspectral</td>
<td>Frequency, ( \forall i \in (1,150) )</td>
</tr>
</tbody>
</table>

### 3.2 Model Performance Metrics

Two primary metrics were used to assess the predictability of the models. The first of these was a standardized version of mean square error (MSE) in order to compare models across response variables since each of the three responses have large differences in the mean by nature of measuring different quantities. To create this standardized MSE (referred to here as SMSE), the MSE was divided by the sum of squares of the true values. The definitions of MSE and SMSE are as follows:

\[
MSE = \frac{1}{n} \sum (y - \hat{y})^2,
\]

\[
SMSE = \frac{\sum (y - \hat{y})^2}{\sum y^2} = \frac{MSE}{\sum y^2}.
\]

Where \( y \) is the true value and \( \hat{y} \) is the prediction of \( y \).
In an effort to report results in a plain manner, the second metric used was the mean absolute percentage error (MAPE) to better understand how much the model deviates from the true value on average in terms of a percentage error.

\[
MAPE = \frac{1}{n} \sum \frac{|y - \hat{y}|}{y}
\]

3.3 Testing and Validation

For each model, 5-fold cross validation was implemented to split the data for training and testing. Both the SMSE and MAPE are calculated for each validation group and the averages were compared to determine the best model.

Reported in Table 4-1, Table 4-2, and Table 4-3 in the results section are the results based on the cross validation split for groups assigned to evenly distribute similar genotypes across each group. As discussed later in the results section, it is evident that there is a significant outlier for Group 5 showing much larger values for the test metrics.

To combat the skewing of the average as a result of this group, testing and validation was also performed using completely random splits into 5 groups as well. The ANOVA analysis, shown previously in Table 2-1, leads us to the conclusion that we can randomly split the data without risk of creating significantly different groups because of genotype. Validation testing was then performed using 5-fold cross validation for 30 iterations with the average metric score reported for each iteration.

4 RESULTS

4.1 Yield prediction

Tested models for predicting yield showed average MAPE under 18%, most of which outperformed the control model with NDVI. All models and metrics by validation group are shown below in Table 4-1. It is worth noting that Groups 1-4 show even lower MAPE values and the average is skewed because of the results from Group 5. Because a few of the yield values in the test split for Group 5 are much lower than the rest of the model, we observe difficulty predicting these values from the training set and a much higher MAPE as the metric also tends to penalize more heavily for predictions greater than the actual value.
### Table 4-1: MAPE Results for yield for each model and validation group

<table>
<thead>
<tr>
<th>Group</th>
<th>Model</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees - PCA</td>
<td>14.04%</td>
<td>11.09%</td>
<td>11.89%</td>
<td>11.61%</td>
<td>39.27%</td>
<td>17.58%</td>
<td></td>
</tr>
<tr>
<td>PCA - Interaction</td>
<td>14.65%</td>
<td>11.47%</td>
<td>11.72%</td>
<td>11.30%</td>
<td>38.90%</td>
<td>17.61%</td>
<td></td>
</tr>
<tr>
<td>PCA - 2nd Order Polynomial &amp; Interaction</td>
<td>14.47%</td>
<td>12.13%</td>
<td>11.72%</td>
<td>11.87%</td>
<td>38.65%</td>
<td>17.77%</td>
<td></td>
</tr>
<tr>
<td>PCA - 3rd Order Polynomial &amp; Interaction</td>
<td>15.25%</td>
<td>11.33%</td>
<td>11.61%</td>
<td>12.03%</td>
<td>38.66%</td>
<td>17.78%</td>
<td></td>
</tr>
<tr>
<td>PCA - 3rd Order Polynomial</td>
<td>15.42%</td>
<td>11.05%</td>
<td>11.53%</td>
<td>12.30%</td>
<td>38.90%</td>
<td>17.84%</td>
<td></td>
</tr>
<tr>
<td>PCA - 2nd Order Polynomial</td>
<td>14.25%</td>
<td>11.14%</td>
<td>11.92%</td>
<td>12.09%</td>
<td>39.80%</td>
<td>17.84%</td>
<td></td>
</tr>
<tr>
<td>Trees - Hyperspectral</td>
<td>13.82%</td>
<td>12.26%</td>
<td>12.64%</td>
<td>11.29%</td>
<td>39.18%</td>
<td>17.84%</td>
<td></td>
</tr>
<tr>
<td>PLS - 1 Component</td>
<td>14.07%</td>
<td>10.74%</td>
<td>12.06%</td>
<td>11.80%</td>
<td>41.09%</td>
<td>17.95%</td>
<td></td>
</tr>
<tr>
<td>Linear – NDVI (Control)</td>
<td>14.03%</td>
<td>10.82%</td>
<td>12.22%</td>
<td>11.85%</td>
<td>40.92%</td>
<td>17.97%</td>
<td></td>
</tr>
<tr>
<td>PCA - 2nd Order Polynomial &amp; Clusters</td>
<td>14.58%</td>
<td>11.57%</td>
<td>11.83%</td>
<td>12.05%</td>
<td>39.93%</td>
<td>17.99%</td>
<td></td>
</tr>
<tr>
<td>PCA - Linear</td>
<td>14.53%</td>
<td>11.59%</td>
<td>11.86%</td>
<td>11.78%</td>
<td>40.33%</td>
<td>18.02%</td>
<td></td>
</tr>
<tr>
<td>PLS - 2 Components</td>
<td>14.50%</td>
<td>11.39%</td>
<td>11.89%</td>
<td>11.99%</td>
<td>40.65%</td>
<td>18.08%</td>
<td></td>
</tr>
<tr>
<td>Clusters</td>
<td>16.50%</td>
<td>11.20%</td>
<td>11.17%</td>
<td>13.89%</td>
<td>40.75%</td>
<td>18.70%</td>
<td></td>
</tr>
<tr>
<td>Ridge</td>
<td>16.95%</td>
<td>12.14%</td>
<td>13.33%</td>
<td>14.47%</td>
<td>40.59%</td>
<td>19.50%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4-1 provides a side-by-side comparison of both model performance metrics using 30 random splits for 5-fold cross validation as described in Section 3.3. The distribution of the 30 averages for each model is show in the figure below. Models are indexed identically for both charts and ordered by descending MAPE on the x-axis. The legend on the right describes which model is in reference by each colored marker. It is evident here that models using PCA tend to perform better than the rest of the models tested. Trees regression with PCA inputs and linear regression using PCA inputs controlling for interaction between the two PCA terms seem to be the two best performers for yield prediction that outperform the NDVI model.
4.2 Gnm2 prediction

Predictive models for Gnm/m² are similar to those of the yield in terms of MAPE. These models performed slightly worse on average but make for decent predictions. Similarly, most PCA tended to outperform other models, including NDVI, as seen in Table 4-2. We observe very similar results for Group 5 as an outlier for this set of validation testing.

Table 4-2: MAPE Results for Gnm2 for each model and validation group

<table>
<thead>
<tr>
<th>Model</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
<th>Group 5</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA - 3rd Order Polynomial &amp; Interaction</td>
<td>17.45%</td>
<td>11.58%</td>
<td>12.37%</td>
<td>12.68%</td>
<td>39.95%</td>
<td>18.80%</td>
</tr>
<tr>
<td>PCA - Interaction</td>
<td>17.35%</td>
<td>11.70%</td>
<td>12.66%</td>
<td>12.25%</td>
<td>40.18%</td>
<td>18.83%</td>
</tr>
<tr>
<td>PCA - 2nd Order Polynomial &amp; Interaction</td>
<td>17.23%</td>
<td>11.56%</td>
<td>12.65%</td>
<td>13.22%</td>
<td>40.34%</td>
<td>19.00%</td>
</tr>
<tr>
<td>PCA - 3rd Order Polynomial</td>
<td>17.90%</td>
<td>11.30%</td>
<td>12.20%</td>
<td>13.49%</td>
<td>40.46%</td>
<td>19.07%</td>
</tr>
<tr>
<td>PLS - 1 Component</td>
<td>17.72%</td>
<td>11.07%</td>
<td>12.48%</td>
<td>12.43%</td>
<td>42.12%</td>
<td>19.16%</td>
</tr>
<tr>
<td>Linear – NDVI (Control)</td>
<td>17.25%</td>
<td>11.27%</td>
<td>12.51%</td>
<td>12.65%</td>
<td>42.76%</td>
<td>19.29%</td>
</tr>
<tr>
<td>PLS - 2 Components</td>
<td>18.31%</td>
<td>11.06%</td>
<td>12.52%</td>
<td>12.74%</td>
<td>42.04%</td>
<td>19.33%</td>
</tr>
<tr>
<td>Trees - Hyperspectral</td>
<td>17.87%</td>
<td>11.31%</td>
<td>13.31%</td>
<td>13.21%</td>
<td>41.15%</td>
<td>19.37%</td>
</tr>
<tr>
<td>PCA - Linear</td>
<td>18.45%</td>
<td>11.10%</td>
<td>12.51%</td>
<td>12.80%</td>
<td>42.31%</td>
<td>19.44%</td>
</tr>
<tr>
<td>Trees - PCA</td>
<td>18.45%</td>
<td>12.09%</td>
<td>13.23%</td>
<td>13.21%</td>
<td>40.28%</td>
<td>19.45%</td>
</tr>
<tr>
<td>PCA - 2nd Order Polynomial</td>
<td>17.75%</td>
<td>11.58%</td>
<td>12.74%</td>
<td>13.99%</td>
<td>41.72%</td>
<td>19.56%</td>
</tr>
<tr>
<td>Clusters</td>
<td>19.42%</td>
<td>14.00%</td>
<td>11.41%</td>
<td>14.05%</td>
<td>40.83%</td>
<td>19.94%</td>
</tr>
<tr>
<td>PCA - 2nd Order Polynomial &amp; Clusters</td>
<td>19.44%</td>
<td>12.71%</td>
<td>12.50%</td>
<td>13.76%</td>
<td>41.96%</td>
<td>20.07%</td>
</tr>
<tr>
<td>Ridge</td>
<td>22.62%</td>
<td>12.00%</td>
<td>15.02%</td>
<td>13.94%</td>
<td>42.84%</td>
<td>21.28%</td>
</tr>
</tbody>
</table>
Table 4-2 shows that models using PCA inputs accounting for interaction between these terms outperformed the other models with very similar results. Looking to Figure 4-2, comparing results based on the 30 sets of 5-fold cross validation, it appears that the model without 2nd and 3rd order polynomial terms outperforms the others despite almost identical results with the initial validation groups.

\[ \text{Figure 4-2: Boxplots of a) MAPE and b) SMSE for \text{Gnm2} by model given 30 random sets for 5-fold cross validation} \]

### 4.3 CMT prediction

Out of the models and responses tested, models predicting CMT are shown to have the best predictive power with only 5% error (Table 4-3). Group 5 has far less influence as an outlier for this response trait. Again, PCA models are shown to have the best predictive power compared to other models tested as evidenced also in \text{Error! Reference source not found.}.
Table 4-3: MAPE Results for CMT for each model and validation group

<table>
<thead>
<tr>
<th>Model</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
<th>Group 5</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees - PCA</td>
<td>4.08%</td>
<td>3.70%</td>
<td>5.45%</td>
<td>4.21%</td>
<td>7.13%</td>
<td>4.91%</td>
</tr>
<tr>
<td>PCA - 2nd Order Polynomial &amp; Clusters</td>
<td>4.11%</td>
<td>4.04%</td>
<td>5.42%</td>
<td>4.54%</td>
<td>7.02%</td>
<td>5.02%</td>
</tr>
<tr>
<td>PCA - 3rd Order Polynomial</td>
<td>4.12%</td>
<td>4.08%</td>
<td>5.55%</td>
<td>4.35%</td>
<td>7.02%</td>
<td>5.02%</td>
</tr>
<tr>
<td>PCA - 3rd Order Polynomial &amp; Interaction</td>
<td>4.25%</td>
<td>4.10%</td>
<td>5.60%</td>
<td>4.36%</td>
<td>7.02%</td>
<td>5.07%</td>
</tr>
<tr>
<td>PCA - Interaction</td>
<td>4.10%</td>
<td>4.02%</td>
<td>5.65%</td>
<td>4.55%</td>
<td>7.16%</td>
<td>5.10%</td>
</tr>
<tr>
<td>PCA - 2nd Order Polynomial</td>
<td>4.13%</td>
<td>4.13%</td>
<td>5.56%</td>
<td>4.67%</td>
<td>7.09%</td>
<td>5.12%</td>
</tr>
<tr>
<td>PCA – Linear (Control)</td>
<td>4.54%</td>
<td>4.11%</td>
<td>5.39%</td>
<td>4.25%</td>
<td>7.43%</td>
<td>5.14%</td>
</tr>
<tr>
<td>PLS - 1 Component</td>
<td>4.57%</td>
<td>4.12%</td>
<td>5.45%</td>
<td>4.31%</td>
<td>7.32%</td>
<td>5.16%</td>
</tr>
<tr>
<td>PCA - 2nd Order Polynomial &amp; Interaction</td>
<td>4.26%</td>
<td>4.16%</td>
<td>5.63%</td>
<td>4.67%</td>
<td>7.14%</td>
<td>5.17%</td>
</tr>
<tr>
<td>Linear - NDVI</td>
<td>4.62%</td>
<td>4.15%</td>
<td>5.31%</td>
<td>4.62%</td>
<td>7.19%</td>
<td>5.18%</td>
</tr>
<tr>
<td>PLS - 2 Components</td>
<td>4.66%</td>
<td>4.10%</td>
<td>5.49%</td>
<td>4.33%</td>
<td>7.43%</td>
<td>5.20%</td>
</tr>
<tr>
<td>Ridge</td>
<td>5.56%</td>
<td>4.36%</td>
<td>5.84%</td>
<td>4.93%</td>
<td>7.64%</td>
<td>5.66%</td>
</tr>
<tr>
<td>Trees - Hyperspectral</td>
<td>6.44%</td>
<td>5.66%</td>
<td>5.41%</td>
<td>4.73%</td>
<td>7.01%</td>
<td>5.85%</td>
</tr>
<tr>
<td>Clusters</td>
<td>5.75%</td>
<td>5.47%</td>
<td>6.68%</td>
<td>5.50%</td>
<td>7.33%</td>
<td>6.15%</td>
</tr>
</tbody>
</table>

Results from the random sets for cross validation show substantial evidence that Trees regression using PCA inputs is a better predictor given that there is little overlap in the variation with the other models as evidenced in Error! Reference source not found.. It is clear that Trees regression using PCA, here, outperforms NDVI as a predictor for CMT.
DISCUSSION & FUTURE WORK

Among the models tested and included in this paper, this represents only a preliminary look into the power of dimension reduction machine learning models to predict yield traits for crops such as grain. To improve upon these models and continue to validate results will require more time-series data and experiments to create more robust models and conclusions. Three points of discussion are outlined below regarding repeatability, weather, and time-considerate models.

5.1 Repeatability

Results outlined previously, for yield and Gn/m² show a large outlier for Group 5. This is because of outliers on the low end for the response in that group as well as difference in the normalized average reflectance curves for Group 5 compared to Groups 1-4, as seen in Figure 5-1.

Because of the variance in MAPE, it is unknown how the introduction of new data would perform with these models. If response traits are more closely normal in distribution without outliers, the models will likely perform better and still be robust for new data, however, many other factors may influence model performance could not be considered such as weather and time that would improve the robustness of the model for new data.

5.2 Weather Considerations
Since the data available included one run from planting to harvest, we are unable to consider weather as an input in the model without better time-series data to compare weather attributes from year to year. Differences in reflectance plots for each of the 4 flights could be explained in part by weather conditions but no trends were found in analysis of major weather indicators. Additionally, given that there are only 4 flights no definitive conclusions can be drawn because of the small sample size.

5.3 Hyperspectral Data Over Time

Given that the models were tested using one set of flights rather than a combination of all flight data it may prove useful to combine the models and methods presented in this paper to create a more robust model that includes time considerations of when the image data is captured. It was evidenced earlier in Figure 2-5 that reflectance curves vary across flights and is likely associated with the changes in the crop as it grows over time. With varying hyperspectral data over time, a model trained previously could be a poor predictor if used too far or too close to harvest date.

6 CONCLUSION

Dimension reduction machine learning algorithms have been shown to be valuable in the modeling of hyperspectral image data to predict yield and related traits. Particularly, the value of PCA has been demonstrated to reduce the high dimensionality of the hyperspectral data while being valuable inputs in supervised learning models such as linear regression and trees regression. Most of the models tested demonstrated strong evidence that they are at least as good as NDVI and even outperform NDVI as a predictor for yield. This is an important result given the quantity of studies that use canopy reflectance measurements and NDVI or other vegetation indices to predict yield.

There is certainly work that can be done to improve these models to consider factors such as weather, time, and multiple sets of flight data; however, this study should prove to demonstrate the value of algorithms such as PCA to leverage technologies such as UAV systems with hyperspectral cameras that provide rapid data collection. In turn, this will enable large improvements in prediction accuracy and crop management practices at low time and labor costs.
7 APPENDIX

7.1 Ridge Regression

Ridge Regression is a similar method to the standard ordinary least squares (OLS) regression. Instead of minimizing the residual sum of squares (RSS), a “shrinkage penalty” is added to the RSS function, shown in the equation below. This shrinkage penalty pushes estimates of the coefficients towards zero. The magnitude of this shrinkage depends on the parameter $\lambda$, which causes the model to behave as OLS when equal to 0 [6].

$$\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \text{Shrinkage Penalty}$$

The model tested in this paper uses a value of $\lambda = 1$ with no substantial changes to report in performance metrics from tuning this parameter.

7.2 Principal Components Analysis (PCA)

PCA is used to reduce the dimensions of a data set. The first principal component that is calculated represents the linear combination of all data points that maximizes its variance. These principal components attempt to explain the variance of the dataset in terms of a few variables rather than the large number of dimensions provided. The second principal component calculated follows the same pattern to find this linear combination using the remaining variables that are uncorrelated with the first principal component. These components are calculated in an unsupervised way, meaning that they do not take into account the response variables. These components can still be used with a regression approach by using the constructed principal components as predictors [6].

As mentioned, the models in this study use the first two principal components only because they were shown to explain over 99% of the variance in the data.

7.3 Partial Least Squares Regression (PLS)

PCA as mentioned in the previous section is an unsupervised learning method used to reduce the dimensions of the dataset and can be used as predictors in a regression model. PLS attempts to improve the PCA regression method by similarly identifying linear combinations of
the original features to create new features that are used as predictors in the regression model. The difference between PLS and PCA is that it uses the response variable to identify these new features that explain the previous features and the response variable [6].

PLS can be executed with any number of components like PCA. The models studied in this paper use models with 1 and 2 components as they were shown to be optimal in the selection of PLS models.

7.4 Trees Regression

Regression trees are used to divide the predictor space into distinct segmented regions. Each observation that falls into the same region will produce the same prediction. Tree-based methods are used in regression and classification applications and are used to address non-linearity in the data. These models operate similarly to forms of human decision making where observations may be sorted based on binary (yes or no) decisions. They receive their name “trees” for the tree diagrams that can be used to visualize these decisions, an example of which is shown in Figure 7-1. The decision splits correspond with regions, shown in Figure 7-2.

Figure 7-1: Tree Diagram Example [6]

Figure 7-2: Predictor Regions Example [6]
REFERENCES


