Paul Hughes

Final Report:
Predictability of Winter Wheat Yield

Problem Statement

The importance of winter wheat to the farmer is twofold: 1) winter wheat is profitable as a cash crop and 2) winter wheat helps build soil fertility as a cover crop. Crop yield measures the quantity of a particular crop produced per unit of land area over a specific period of time, and is a fundamental factor that influences the overall profitability of a farm. Therefore, is it possible to predict the yield of winter wheat to help mitigate the farmer's risk and maximize their profitability?

Winter wheat is typically planted in the fall season and is harvested in the summer season of the following year. In contrast, corn, soybeans, cotton, and rice are typically planted in the spring and harvested in the fall of the same year. Therefore, this work seeks to find if the yield of corn, soybeans, cotton, and rice can be used as predictors for the yield of winter wheat. For example, can the 2023 yield of corn, soybeans, cotton, and rice predict the 2024 yield of winter wheat?

To understand the initial degree of predictability of winter wheat, this work strictly used crop yield data for the state of Missouri and explored two regression techniques: 1) multiple linear regression and 2) random forest regression.

Data Wrangling

The raw data set from the United States Department of Agriculture (USDA) contained 365 rows with 21 columns. The rows represent the yearly yield of corn, cotton, rice, soybeans, and winter wheat for 1951–2023 in the state of Missouri. On inspection, it was determined that many of the columns did not contain useful information (17 in total), so they were immediately eliminated from the dataset. The trimmed dataset was reduced to four columns, which provided information about the specific year, the name of the commodity (e.g., corn, rice, soybeans, etc.), the units of measurement, and the yield value, respectively. To clean the data further, the crop yield values were modified to eliminate the presence of commas and, then, converted from strings to floating points. At this point, the dataset was deconstructed into five separate Pandas DataFrames (i.e., one dataframe for each of five commodities). This deconstruction allowed for the individual DataFrames to be concatenated so that each row of the new dataset represented a particular year and the columns contained the yield data for each of the five commodities: corn, cotton, rice, soybeans, and winter wheat. However, prior to concatenation, the yield data for cotton and rice was converted from pounds per acre to bushels per acre using 32 pounds per bushel and 45 pounds per bushel as the conversion rates, respectively. In the end, the final dataset contained 72 rows and 7 columns.

Exploratory Data Analysis

The viability of predicting this year's winter wheat yield from the previous year's yield of corn, cotton, rice, and soybean depends on whether a relationship exists between the yield of winter wheat and that of the other crops. Figure 1 shows that a positive linear relationship exists between the yield of winter wheat and the previous year's yield of corn, cotton, rice, and soybean, respectively. This means that lower yields of corn, cotton, rice, and soybean correspond to lower yields of winter wheat. Similarly, higher yields of corn, cotton, rice, and soybean correspond to higher yields of winter wheat. Based on these results, the prediction of this year's winter wheat wheat yield from the previous year's yield of corn, cotton, rice, and soybean appeared viable.

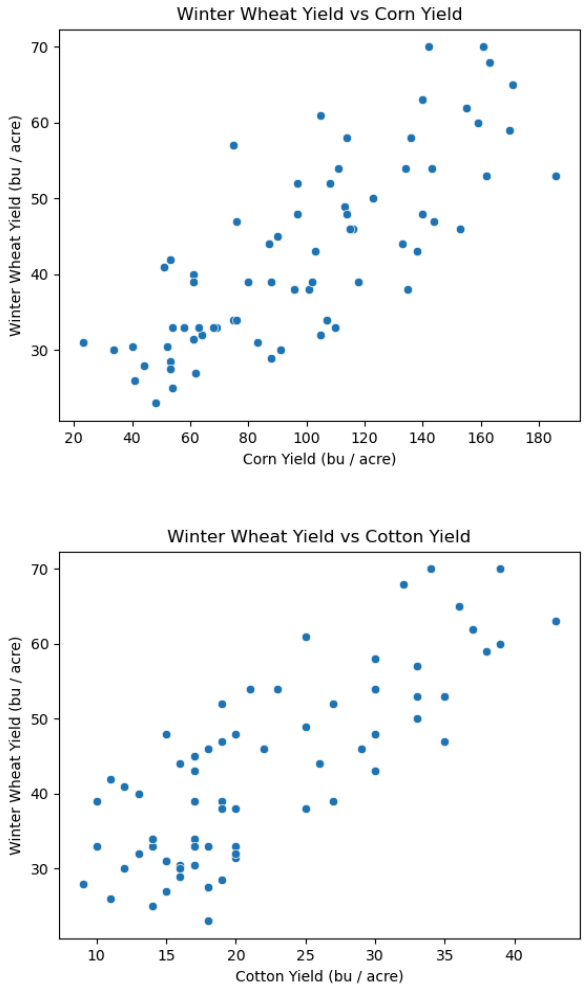

Figure 1. Scatterplots of the winter wheat yield versus the previous year's yield of corn (top) and cotton (bottom).

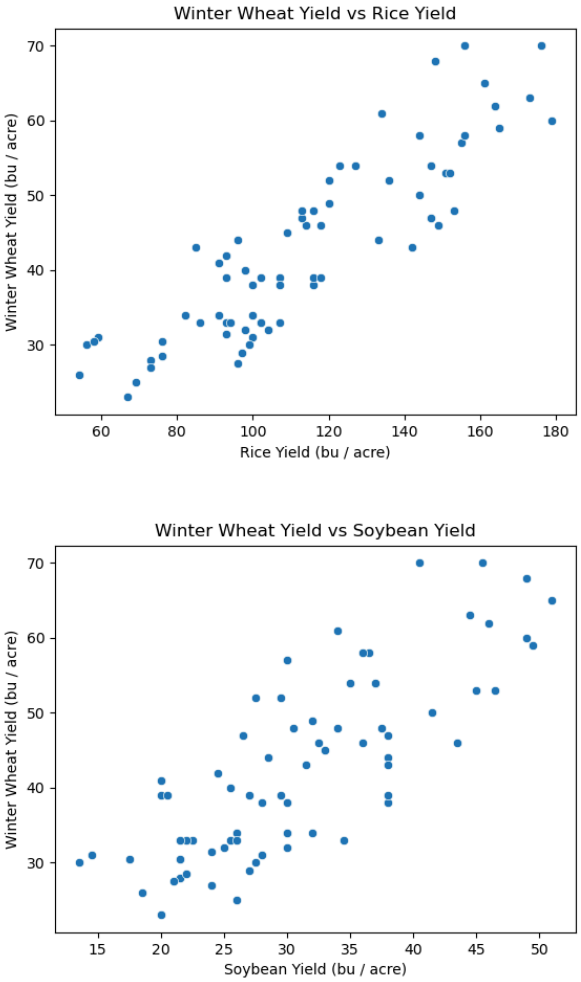

Figure 1 continued. As in Figure 1 except for rice (top) and soybean (bottom).

Modeling Results

The predictability of the winter wheat yield was explored using two models: 1) linear regression and 2) random forest regression. Prior to applying the models, the data associated with the features (i.e, the yields of corn, cotton, rice, and soybean) was scaled and, then, the whole dataset was divided into a test and training set for model evaluation.

Linear regression

In conjunction, Figure 2 and Table 1 summarize the modeling results compared to ground truth for the training and test datasets, respectively. Overall, the linear regression model performs well. For both the training and test datasets, the linear regression model has root mean squared errors (RMSE) and mean absolute errors (MAE) between 4–6 bushels per acre. In addition, the R-squared values indicate that approximately 76% of the variance in the winter wheat yield is explained by previous year's yield of corn, cotton, rice, and soybean. The cross validation results suggest the linear regression model is stable with respect to new data.

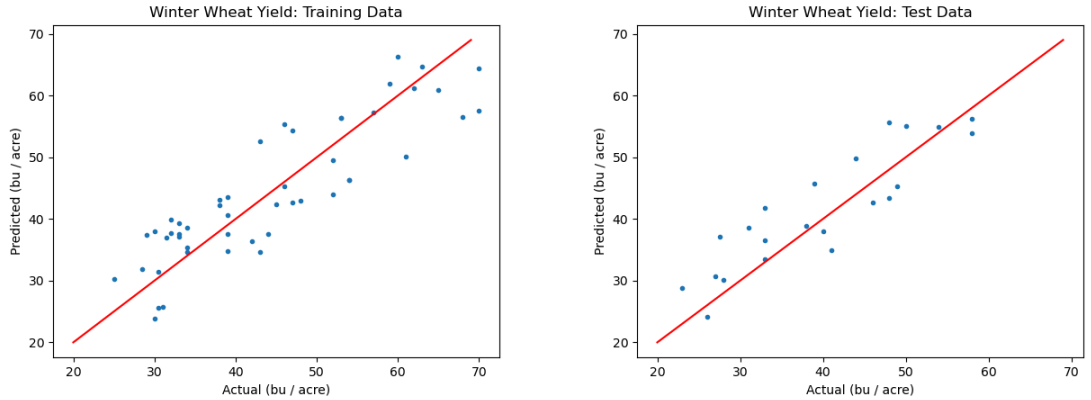

Figure 2. Scatterplots of the predicted values of the winter wheat yield compared to ground truth for the training dataset (left) and test dataset (right) using a linear regression model. The red line represents a perfect one-to-one relationship.

Table 1. Evaluation of linear regression model to ground truth.

	Training Data	**Test Data**
Mean Squared Error (MSE)	35.10	25.73
Root Mean Squared Error (RMSE)	5.92	5.07
Mean Absolute Error (MAE)	5.15	4.37
R-squared	0.77	0.76
Mean 5-fold cross validation score	0.72	—-
Standard deviation of cross validation score	0.08	—-

Random forest regression

Together, Figure 3 and Table 2 summarize the modeling results compared to ground truth for the training and test datasets, respectively. Overall, the random forest model for the training data explains approximately 96% of the variance in winter wheat yield and has a RMSE and MAE around 2 bushels per acre. Thus, for the training data set, the random forest regression model performs better than the linear regression model. However, for the test dataset, the random forest model does not show this marked improvement over the linear regression model. Similar to the linear regression model, for the training dataset, the random forest model has root mean squared errors (RMSE) and mean absolute errors (MAE) between 4–6 bushels per acre and a R-squared value of 0.76. Regarding the cross validation results, the random forest model shows more variability with respect to new data than the linear regression model.

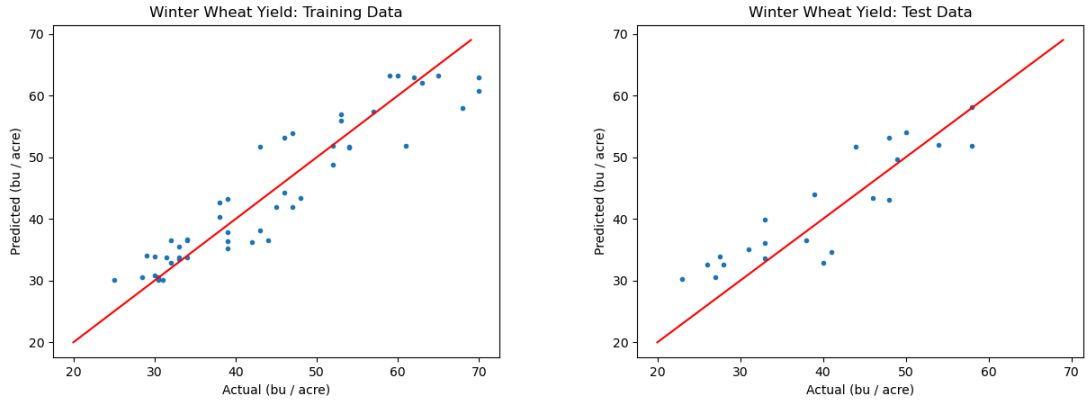

Figure 3. Scatterplots of the predicted values of the winter wheat yield compared to ground truth for the training dataset (left) and test dataset (right) using a random forest regression model. The red line represents a perfect one-to-one relationship.

Table 2. Evaluation of random forest regression model to ground truth.

	Training Data	**Test Data**
Mean Squared Error (MSE)	6.00	26.19
Root Mean Squared Error (RMSE)	2.45	5.12
Mean Absolute Error (MAE)	2.03	4.52
R-squared	0.96	0.76
Mean 5-fold cross validation score	0.67	—-
Standard deviation of cross validation score	0.12	—-

Further research

With respect to the test data, the linear regression and random forest model performed similarly in predicting the winter wheat yield. Because the linear regression model had a slightly better RMSE, MAE, and mean 5-fold cross validation score, it is the model that will be used going forward. To improve upon the current linear regression model, there are currently three main areas for further research:

1. Including crop yield data from neighboring states (e.g., Kansas, Nebraska, Iowa, and Illinois)

2. Adding information through modes of climate variability (e.g., El Nino-SouthernOscillation, Pacific Decadal Oscillation, Arctic Oscillation, North Atlantic Oscillation, etc.)

3. Adding information through soil information prior to winter wheat planting

4. Investigate why the linear regression model appears to overestimate low yields of winter wheat.

Client Recommendations

This study shows that it is possible to accurately predict the yield of winter wheat in Missouri given the previous year's yield of corn, cotton, rice, and soybean, helping to mitigate the farmer's risk and maximize their profitability. However, at present, the model is not ready for real-time application. It is hypothesized that the linear regression model's performance can be improved through the inclusion of additional crop yield data, climate data, and/or soil data.