

**United States Department of Agriculture
Project Initiation**

Project Data for Accession Number 1030766

Project Title

Using Machine Learning and Geospatial Data to Generate Near Real Time Crop Yield and Production Forecasts.

Project Details

Sponsoring Institution National Institute of Food and Agriculture	Project Status ACTIVE
Funding Source NIFA Non Formula	Grant Year 2023
Grant No. 2023-67023-40219	Cumulative Award Amt. \$429,333.00
Proposal No. 2022-10664	Multistate No. (N/A)
Project Start Date May 15, 2023	Project End Date May 14, 2027
Program Code [A1641] Agriculture Economics and Rural Communities: Markets and Trade	

Project Director

Mishra, A.

Recipient Organization

ARIZONA STATE UNIVERSITY
660 S MILL AVENUE STE 204
Tempe, ARIZONA 852813670

Performing Department

(N/A)

Non Technical Summary

This project aims to reduce the potential information gap between large traders (e.g., institutional traders; grain merchandisers; large agribusinesses) and small / medium sized farms through the deployment of near real-time forecasts of crop yield and production for a set of grain commodities. Thus, the primary activities of this project are twofold. First, assess the accuracy of crop yield and production forecasts made using a machine learning algorithm, XGBoost, with satellite and weather data, and compare their accuracy to the WASDE yield and production forecasts. Second, create an automated forecasting system that generates daily yield and production forecasts based on this machine learning algorithm, with the estimates released to the public through an automated report distributed through a web portal. Indeed, many institutional traders, grain merchandisers, and other large agribusinesses pay large sums for private forecasts of key production variables. These private forecasts can be prohibitively expensive for smaller and medium-sized farms and agribusinesses, resulting in an information gap or information disadvantage to these

United States Department of Agriculture Project Initiation

smaller businesses. A low cost, widely distributed, informationally efficient forecast utilizing machine learning algorithms may benefit producers and market participants along the grain supply chain. Therefore, the proposed activities of both testing and deploying near real-time machine-learning based forecasts of yield and production, at multiple levels of aggregation, will help producers better anticipate crop market prices and market supply. Furthermore, this research will help broaden our understanding of the efficacy of machine learning and big data approaches to forecasting key production variables important for agriculture.

Animal Health Component

Animal Health Component: 0%

Research Effort Categories

Basic	0%
Applied	100%
Developmental	0%

Classification

Knowledge Area (KA)	Subject of Investigation (SOI)	Field of Science (FOS)	Percent
603	6110	3010	100%

Knowledge Area

[603] Market Economics

Subject Of Investigation

[6110] Economy of the United States and sectors thereof

Field Of Science

[3010] Economics

Keywords

Decision making, Forecasting, Crop Yield, Big data, Market efficiency

Goals / Objectives

Agricultural forecasting models are essential for assisting agricultural producers, consumers, governments, crop insurers, and other market participants to more effectively anticipate crop yield and supply (Allen, 1994). Crop yield and production forecasts help the global supply chain become more efficient in meeting food demand and improving food security (Ahumada et al., 2016). Crop yield forecasts are also essential to many market participants and provide informational value for commodity markets (Crespo et al., 2021) and may also be a valuable input for price forecasting (Ahumada et al., 2016; Crespo et al., 2021, Crespo et al., 2018; Gargano and Timmermann, 2014). The U.S. Department of Agriculture (USDA) has two of thirteen federal government forecasting agencies: the National Agricultural Statistics Service (NASS) and the Economic Research Service (ERS). These agencies account for approximately 8% of the \$3.2 billion annual budget allocated to federal forecasting agencies (Bora, Katchova, and

United States Department of Agriculture Project Initiation

Kueth, 2020). These agencies release several forecast reports including the World Agricultural Supply and Demand Estimates (WASDE) report. WASDE is an essential monthly forecast report prepared and released by the USDA's World Agricultural Outlook Board (WAOB). The report contains information on agricultural production, consumption, inventories, trade, and prices for a number of key agricultural commodities. The industry closely follows these reports, and there is evidence that commodity markets respond to WASDE estimates (Bora, Katchova and Kueth 2020; Summer and Mueller 1989; Fortenberry and Sumner 1993; Isengildina-Massa et al. 2008; McKenzie 2008; Dorfman and Karali 2015; Adjemian and Irwin 2018; Karali et al. 2019; Isengildina-Massa et al. 2021). NASS crop yield and production forecasts for corn and soybeans are released in the WASDE report from August to November. Agency forecasts are generally created using sophisticated sampling and survey methodologies. For example, for state crop yield forecasting, NASS uses data collected from the Average Yield Survey (AYS) and the Objective Yield Survey (OY). Then forecasts of state yields are weighted to the national level through a consensus estimating process to create the monthly national corn yield forecast released in the WASDE report (USDA NASS, 2012; USDA NASS and WAOB, 1999). Satellite imagery and weather data are used by the Foreign Agricultural Service (FAS) for estimating crop yields and production for foreign countries, but not for domestic forecasting. There are several studies that have investigated adopting machine learning forecasting methods and using large geospatial data such as weather, soil, and satellite data. Some notable examples include Kang et al., 2020, Peng et al., 2018, and You et al., 2017. Also, government forecasting agencies in other countries have adopted these approaches (Cerrani and López Lozano, 2017; Chipanshi et al., 2015). Machine learning models may produce reasonably accurate crop yield forecasts and have two main advantages. First, forecasts can be made at high frequency (e.g., daily), and second, generating forecasts is less resource intensive than traditional survey-based methods. Once a machine learning model is developed, forecasts can be generated automatically with little human involvement and be presented in a web portal or through an automated report. This can reduce the cost of creating forecasts and allows for high-frequency forecasting that can adapt to current growing season conditions. In contrast, survey-based methods require significant resources including skilled staff to collect data, prepare analysis, interpret results, and form a consensus. Also, survey methods rely on the responses of farmers, which must be collected during busy times in the growing season (e.g., planting and harvest). Satellite and weather data are globally available and provide near-continuous data (e.g., daily data) that can be used for crop yield forecasting in important agricultural growing regions. The quality of this data is continuously improving as satellite technology evolves and as data history increases. Machine learning models utilizing satellite imagery and weather data may help enhance and augment traditional resource-intensive survey-based crop yield forecasts. Machine learning methods have improved rapidly in the past decade, and many algorithms have emerged which are well suited for extensive complex big data such as satellite images and weather data. Despite the advantages satellite data and weather data pose for agricultural forecasting, widespread adoption has not been achieved. A limitation of agricultural forecasting using weather data and satellite image data is that significant effort is required to prepare the data for modeling. Processing is time-consuming, computationally intensive, and requires expert knowledge. Whether machine learning or traditional statistical approaches are applied, the weather and satellite data must be cleaned and aggregated. Machine learning-generated agricultural forecasts may be valuable for producers and market participants because they can be produced at a high frequency (e.g., daily) compared to survey forecasts (monthly). For example, producers may be able to better anticipate the effect of weather events ahead of monthly government forecast releases and use the information for farm-level decision-making. Many institutional traders, grain merchants, and other large agribusinesses pay considerable sums for private forecasts, including machine learning-based forecasts. This information can be prohibitively expensive for small and medium-sized businesses, in particular small farms, and may result in an information gap between larger, more capitalized businesses and producers. This project aims to reduce the informational gap between large market participants (e.g., institutional traders; grain merchandisers; large agribusinesses) and small / medium-sized farms through the deployment of near real-time forecasts of crop yield and production incorporating machine learning. Therefore, the specific objectives of this study are twofold. First, focusing on both corn and soybeans, we will assess the accuracy of crop yield and production forecasts made using a machine learning algorithm, XGBoost, incorporating both satellite and weather data, and compare their accuracy to the WASDE yield and production forecasts. Competing machine learning algorithms such as neural network models, and regression trees will be investigated. Our focus on corn and

United States Department of Agriculture Project Initiation

soybeans for this study is important as numerous market advisory, brokerage, and other trading firms provide their own (private) forecasts for corn and soybean yield and production often for hefty fees. Second, we will create an automated forecasting system that generates daily yield and production forecasts for corn and soybeans utilizing the tested machine learning algorithm. The automated estimates will be released to the public through an automated report distributed through a web portal. Sharing these forecasts broadly with market participants may reduce the information gap between small / medium producers and more capitalized agribusinesses, leading to greater market efficiency and better farm decision-making. Furthermore, the machine learning algorithm used efficiently synthesizes disparate publicly available information from many sources, thus providing a mechanism for this data to be used by market participants across the grain marketing supply chain. Indeed, these improvements are likely to benefit producers, market participants, and consumers alike.

Project Methods

For decades, government agencies and private companies have collected satellite images and weather data. Many government forecasting agencies and industry forecasters use this data to forecast crop yields and production. Due to the large size and complexity, governments and well-funded research institutions have historically been the primary users of the data. Many studies have developed models for forecasting in-season crop yields at various levels of aggregation from the field level to the national level (Cerrani and López Lozano, 2017; Chipanshi et al., 2015; Johnson et al., 2021; Paudel et al., 2021; Paudel et al., 2022). Satellite images and weather data are generally large (in other words, big data), making processing and preparing for modeling complex (Liu, 2015). However, technological advances and specific improvements in cloud computing have made the data more accessible to a larger audience of forecasters. In addition to reductions in cost and improved accessibility, modeling techniques and algorithms that handle big data have vastly improved. At the time of writing, satellite and weather-based models can be created for a relatively low cost using free and open source tools, including Python and Google Earth Engine (Gorelick et al. 2017). As technology advances, developing sophisticated models is expected to become less expensive and easier (Ma et al., 2015). Over the past decade, machine learning methods have advanced and been driven by new challenges posited by the vast expansion of data. In partnership with the research community, companies like Facebook, Google, and Amazon have developed and applied new machine learning algorithms and statistical tools to extract insights from large data sets. Some examples include XGBoost, LightGBM, Prophet, Support Vector Machines, and Deep Neural Networks (Crane-Droesch, 2018; Tedesco-Oliveira et al., 2020; Van Klompenburg et al., 2020). These machine learning methods are well suited to work with large and complex data such as satellite images and weather data. There may be opportunities to incorporate these methods and improve existing government statistical forecasting methods. Many factors determine end-of-season crop yield, and a simple theoretical framework may include weather, managerial, technological, and location factors. In the case of in-season forecasts, as discussed in this proposal, variables are measured up to the forecast date and used to predict the end-of-season crop yield as reported in WASDE. Forecasts can be made at multiple levels of aggregation from the pixel level (fraction of a field) to the national level. In the case of national-level forecasts, forecasters may choose to forecast national-level yields directly or use a hierarchical approach, in which forecasts made at lower levels of aggregation are weighted to the national level. This is the approach considered in this proposal and discussed below. An empirical crop yield forecast model is developed that 1) generates crop yield forecasts at the state level and then 2) aggregates those forecasts to the national level. The machine learning algorithm, XGBoost, is used in step one to generate state-level forecasts. A linear regression model with national-level yields regressed on state yields is used in step two to weight the XGBoost forecasts and aggregate to the national level. Also, the framework is adjusted for county-level yield and production forecasts. The state corn yield model is outlined below for ease of illustration. Corn yields are forecasted at the state level using monthly normalized difference vegetation index (NDVI) values, accumulated precipitation, accumulated maximum and minimum vapor pressure deficit, and a technology trend. The theoretical framework outlined in Equation 1 describes crop yields as a function of weather, managerial, technological, and location factors. The variables NDVI, accumulated precipitation, accumulated maximum and minimum vapor pressure deficit, and growing degree days are included in the model to capture weather factors and interactions with managerial factors and location factors. Weather data are used in addition to satellite NDVI because NDVI measurements could indicate high greenness and

United States Department of Agriculture Project Initiation

high yield. Still, very high temperatures or very low temperatures could reduce yield. For example, very high temperatures may reduce seed filling, decreasing yield but having little to no effect on greenness or biomass. NDVI may also capture the effects of management and location effects on crop yields. For example, higher quality soils may result in higher NDVI. Equation 2 shows the generalized empirical state corn yield model. State corn yield forecasts are generated from the XGBoost model and then aggregated to the national level. The method used is called Extreme Gradient Boosting (XGBoost), which is a tree-based boosting machine learning algorithm. XGBoost was introduced in 2016 and has become a popular machine learning algorithm for regression and classification problems. XGBoost is an extension of earlier gradient boosting algorithms developed by Friedman in 2002, which improves the scalability of the gradient boosting algorithm (Chen and Guestrin, 2016). Following the notation of Friedman (2002) gradient boosting is used to solve a function estimation problem by mapping a set of known response variables and explanatory variables. The state forecasts generated in step one are then aggregated to the national level. First, a model is specified, shown in Equation 4, to weight the state-level yields to the national level. This model is estimated using feasible generalized least squares (FGLS) with 15 years (or another predefined window length) of historical state-level yields from the top 10 producing states. Once the model is estimated, the state-level forecasts generated using XGBoost are substituted for the actual state yields to create national-level forecasts. The data are split into training and testing data sets to avoid model overfitting and to replicate an actual forecast using only data available to the forecaster at the forecast date. Seventy percent of the data are designated as the training set and thirty percent as the testing set. Initial model development, including variable development, variable selection, functional form specification, and regression diagnostics, is performed on the training set. Once the model development is complete, the model is estimated using a walk forward hold out forecast process, re-estimating the model at each walk forward step. For example, for the 2021 forecast, the state-level XGBoost model is estimated using data from the earliest record date to 2020, and state forecasts are generated from the estimated model for 2021. The national level weighting model is then estimated using data from a predefined window length (e.g., 15 years). The state forecasts for 2021 are then substituted into the national-level weighting model to generate the 2021 national-level forecast. Once the national forecasts are generated, they are compared relative to the NASS released actual yield estimates (released in January after the growing season and reported in the corresponding WASDE report) and relative to the WASDE corn yield forecasts and a Naïve forecast. Three national forecasting methods are compared. WASDE forecast - created by NASS using a survey methodology and aggregated from state-level to the national level by a consensus committee. XGboost forecast - a machine learning forecast approach that generates state-level corn yield forecasts using satellite images and weather data collected up to the forecast date. The state forecasts are then weighted to a national forecast by a linear model estimated with actual historical state-level and national-level yield data. Naïve forecast - a national level forecast using the corn yield from last year as the forecast.