In-season Sweetpotato Yield Forecasting using Multitemporal Remote Sensing Environmental Observations and Machine Learning

Mariella Carbajal-Carrasco[1,2,3], Daniela Jones[1,2,4,5], Cranos Williams[2,3], and Natalie Nelson[1,2,4]

[1]Biological and Agricultural Engineering Department, North Carolina State University
[2]N.C. Plant Sciences Initiative, North Carolina State University
[3]Electrical and Computer Engineering Department, North Carolina State University
[4]Center for Geospatial Analytics, North Carolina State University
[5]Operations Research and Analysis Group, Idaho National Laboratory

April 26, 2024

This is a preprint of a manuscript submitted for peer review.

In-season Sweetpotato Yield Forecasting using Multitemporal Remote Sensing Environmental Observations and Machine Learning

Mariella Carbajal-Carrasco[a,b,e], Daniela Jones[a,b,c,d], Cranos Williams[b,e], Natalie Nelson[a,b,c,*]

[a]*Biological and Agricultural Engineering Department, North Carolina State University, Raleigh, NC*
[b]*N.C. Plant Sciences Initiative, North Carolina State University, Raleigh, NC*
[c]*Center for Geospatial Analytics, North Carolina State University, Raleigh, NC*
[d]*Operations Research and Analysis Group, Idaho National Laboratory, Idaho Falls, ID*
[e]*Electrical and Computer Engineering Department, North Carolina State University, Raleigh, NC*

Abstract

Data-driven modeling approaches for crop yield prediction have exponentially increased in the last decade due to the greater availability of spatial data from various sensors. Yet, most yield modeling has focused on major commodities, leaving lesser-cultivated horticultural crops like sweetpotato relatively undertooled, though these crops considerably contribute to the global economy and food supply. The U.S. is the primary exporter of sweetpotato (271 K tonnes), with 21% of U.S.-grown sweetpotatoes being exported. Early yield forecasting at the county scale offers crucial insights for growers, packers, wholesalers, and associated industries, enabling them to anticipate variations in yield to make informed decisions. While roots and tubers have demonstrated a relationship between yields and above-ground plant characteristics, it remains uncertain whether forecasting models that utilize remotely sensed data, including vegetation indices, are suitable for sweetpotato. We developed county-scale in-season sweetpotato yield forecast models using machine learning (ML) algorithms and multitemporal remote sensing environmental data. Four of the most commonly used ML algorithms for predicting crop yield - Random Forest Regression (RFR), Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB) - were applied using stationary (topography and soil characteristics), and temporal (weather, NDVI, and Growing Degree Days) variables as potential predictors. Six predictor sets were tested to identify key predictor variables, optimal aggregation time (16 or 32 days composite) of the temporal variables, and how early in the growing season the models can reliably predict end-of-season yields. U.S. Annual CropScape land cover layers were used to identify sweetpotato fields, over which temporal variables were aggregated, and sweetpotato yields were tabulated from the USDA Agricultural Survey from 2008 to 2022. The Boruta method was used for feature selection across each predictor set before training the ML models. RFR outperformed other ML algorithms and the RFR models' evaluation metrics were the most consistent across the six predictor sets. The RFR model that incorporated early and mid season temporal variables as 16-day composites was selected and proposed for future sweetpotato yield forecasting due to its performance ($R^2 = 0.44$, RMSE = 3.53 $tonnes.ha^{-1}$), as well as ability to predict early enough

in the season to provide actionable information. In the final model, several stationary variables (elevation, nitrogen, cec, soc, and clay content) were the most predictive of sweetpotato yield. After these stationary variables, NDVI and precipitation from the time around storage root initiation and bulking (July), and minimum temperature around planting (June) followed in importance.

Keywords: Random forest, XGBoost, neural networks, yield prediction model, vegetation index, NDVI, Sentinel-2, Landsat

1. Introduction

Innovations in technology and access to remote sensing data have driven huge advances in the application of artificial intelligence to agriculture, such as to predict yields (Jung et al., 2021). In particular, machine learning (ML) algorithms are effective at crop yield prediction due, in part, to their ability to use observational data and measurements across several experiments. Additionally, non-linear ML algorithms do not assume a defined pattern (i.e. linear, polynomial) between the predictor and response variables and can account for non-linear relationships evident through patterns in recorded data (Paudel et al., 2022), making them well suited for use with agricultural observations. Even though ML models are unable to explain underlying processes, they can surpass the predictive accuracy of process-based models (Leng and Hall, 2020), making ML algorithms particularly useful for yield forecasting at scales that are often computationally prohibitive for process-based models. Artificial Neural Networks and Random Forest Regression are among the most used and successful ML algorithms for predicting crop yield, specifically using agroclimatic variables derived from remote sensing products as predictors (Van Klompenburg et al., 2020).

While ML-based yield forecasting has become increasingly common, most yield forecasting research has predominantly concentrated on field-scale predictions (Van Klompenburg et al., 2020). These forecasts are highly site-specific and less suitable for extrapolation to other fields. A few studies have put forth crop yield prediction models at the county scale, but only for major crops - e.g., wheat, corn and soybean (Cao et al., 2021; Zhou et al., 2022; Kang et al., 2020; Ghazaryan et al., 2020) - leaving growers of non-major commodities without forecast information to inform farm management. Moreover, it is unclear whether existing forecasting frameworks that work effectively for crops like wheat, corn and soybean will transfer to other crops with distinct physiology, like roots and tubers. Although roots and tubers grow underground, studies suggest a relationship between yield and above-ground traits, such as vegetation indices or canopy cover, particularly during vegetative growth and around tuber or storage root initiation (Tedesco et al., 2021; Pérez-Pazos et al., 2021;

*Corresponding author
Email address: nnelson4@ncsu.edu (Natalie Nelson)

Sun et al., 2020). The relationship between yields and above-ground plant characteristics indicates that remotely sensed data may still be useful when predicting root and tuber yield.

Among yield forecasting studies focused on roots and tubers, greater attention has been dedicated to potato. Prior potato forecasting studies have demonstrated that high-resolution vegetation bands and indices (from UAV or Sentinel-2) acquired during full vegetative growth and around tuber initiation were predictive of potato yield, at both field (Sun et al., 2020; Gómez et al., 2019) and regional scales (Salvador et al., 2020). Additionally, tuber set was better predicted than tuber yield due to its higher correlation with above-ground biomass monitored with spectral data (Sun et al., 2020). Furthermore, studies that incorporated more precise in-situ data, such as cultivar information (e.g., plant height) (Li et al., 2021) or soil characteristics, along with proximal data (Abbas et al., 2020) during the growing season, achieved higher prediction performance.

In contrast to potato, ML-based forecasting models have not yet been tested or developed for sweetpotato. The complex interplay of multiple genotype traits, phenological dynamics, and environmental factors in sweetpotatoes across tropical and some temperate regions, along with limited availability of in-season data, poses challenges for accurately predicting sweetpotato yield. Despite the lack of attention sweetpotato has received from yield forecasters, sweetpotato is a key crop in regions of the United States of America (USA), which is the primary exporter of sweetpotato globally. In 2021, the USA exported 271 K tonnes of sweetpotatoes (USDA National Agricultural Statistics Service, 2022b). Within the USA, North Carolina (NC) is the largest sweetpotato-producing state, generating nearly half of the national sweetpotato supply and having influence on national (Soto-Caro et al., 2022) and international markets. Having access to a sweetpotato forecasting model would help growers, packers, wholesalers, and supporting businesses (e.g., exporters, crop insurers) have information with which to anticipate and respond to yield deviations.

While a sweetpotato forecasting model does not yet exist, prior studies have tested ML algorithms for predicting other aspects of sweetpotato production. Villordon et al. (2009a) evaluated eight growing degree day (GDD) calculation methods with three base temperatures (60, 65 and 70 F), five ceiling temperatures (80, 85, 90, 95, 100 F), and six machine learning algorithms (support vector machine, multivariate adaptative regression, neural networks, linear regression, regression tree, and generalized linear model) to identify suitable models for predicting optimal harvest dates. Then, Villordon et al. (2010) used a Bayesian Belief Network (BBN) approach to identify the agroclimatic variables known to influence critical storage root initiation in marketable sweetpotato yield. Similarly, Villordon et al. (2011) used a BBN and data on agroclimatic conditions to determine the optimal in-row spacing to reach the highest marketable yield. Combined, these studies demonstrate the importance of optimum air and soil temperatures during storage root formation, insights on which can be used towards developing a sweetpotato yield forecasting model.

In this study, we developed county scale in-season sweetpotato yield forecast models using ML algorithms and multi-temporal remote sensing, and environmental data. Specifically, our objectives were to (1) identify key predictor variables through feature selection from candidate variables including Normalized Difference Vegetation Index (NDVI), maximum temperature, minimum temperature, precipitation, GDD, topography, and soil properties, (2) implement four ML algorithms, specifically Random Forest Regression (RFR), Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB), to predict in-season sweetpotato yields at the county scale, (3) determine the optimal aggregation periods for temporal predictor variables, and (4) evaluate how early in the growing season models are able to reliably predict end-of-season yields. We focused on North Carolina (NC; USA) counties in and around the Coastal Plain agro-region, where the highest sweetpotato production occurs in the state.

2. Materials and methods

2.1. Study area

We focused on NC counties with reported sweetpotato production (USDA National Agricultural Statistics Service, 2022b) from 2008 to 2022 (Figure 1), which comprises major sweetpotato producers and exporters. The majority of the counties were located in the Rolling Coastal Plains (Level IV) and Southeastern Plains (Level III) ecoregions (Griffith et al., 2002), where environmental conditions are ideal for sweetpotato growth. All counties for which yield data (ranging between 12.55 to 33.89 $t.ha^{-1}$) and predictor variables (n = 17) were available were considered in this analysis. These counties are outlined in Figure 1 and included: Johnston, Sampson, Chowan, Edgecombe, Harnett, Nash, Wake, Wayne, Wilson, Robeson, Duplin, Lenoir, Pitt, Columbus, Martin, Lee, and, Moore. To narrow down the location of sweetpotato fields within each county, counties were also matched with gridded sweetpotato areas classified by the Cropland Data Layer (CDL, 30-m resolution), hosted in the web-based tool CropScape (USDA National Agricultural Statistics Service, 2022a). CDL was used to identify sweetpotato harvested areas every year (pixel code: 46). Figure 2 depicts how sweetpotato harvested areas nearly doubled from 2008 (120 km^2) to 2022 (229 km^2). Figure 2 also illustrates the concentration of sweetpotato fields within the studied counties.

2.2. Modeling approach

Crop yield is driven by interactions between genetics, environment, and management (Gajanayake et al., 2014). When developing a crop yield forecasting model for use across a region, only environmental variation can be directly accounted for, as genetics and management will vary by farm. While genetic and management data cannot be included in a model designed for regional application, the

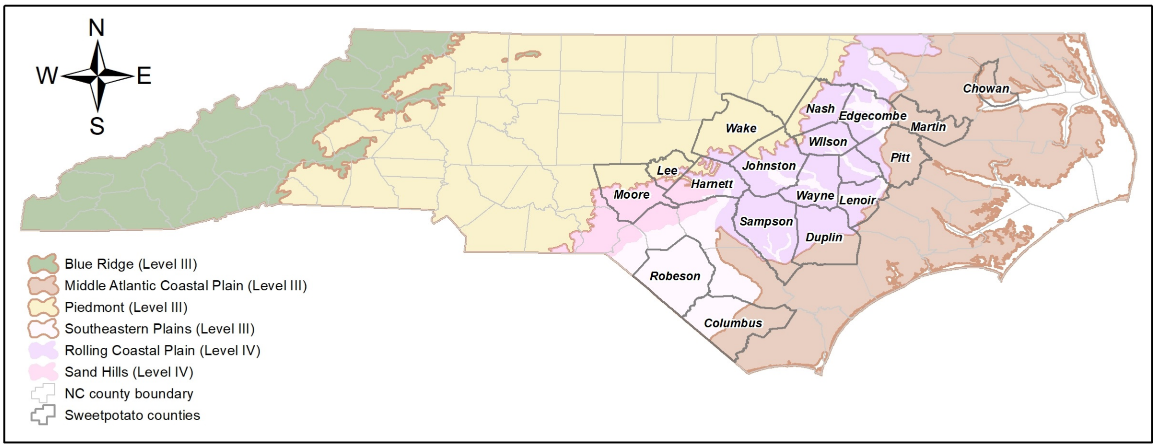
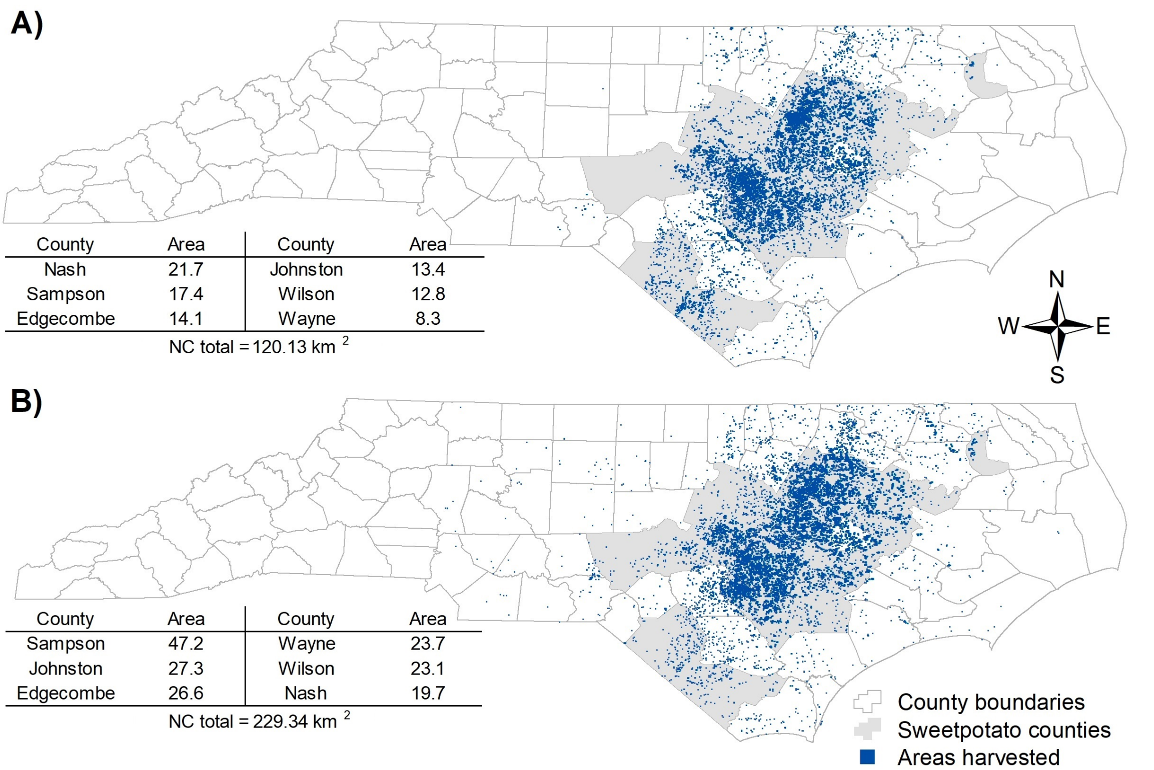


Figure 1: North Carolina counties that reported sweetpotato harvested areas and yield in USDA Statistics (USDA National Agricultural Statistics Service, 2022b) from 2008 to 2022



A)

County	Area	County	Area
Nash	21.7	Johnston	13.4
Sampson	17.4	Wilson	12.8
Edgecombe	14.1	Wayne	8.3

NC total = 120.13 km ^{2}

B)

County	Area	County	Area
Sampson	47.2	Wayne	23.7
Johnston	27.3	Wilson	23.1
Edgecombe	26.6	Nash	19.7

NC total = 229.34 km ^{2}

Figure 2: Sweetpotato harvested areas (km^2) in A) 2008 and B) 2022 derived from the Cropland Data Layer (CDL) (USDA National Agricultural Statistics Service, 2022a). Counties with gray fill have reported sweetpotato yields from USDA NASS. Colored pixels show sweetpotato fields reported in the CDL; pixels are magnified for visualization purposes. The table in each panel summarizes the CDL top six counties in terms of areas harvested, as well as the total harvested area across NC

characteristics of common genotypes and management practices can inform the selection and summary of candidate predictor variables. Here, we considered both stationary (i.e., topography and soil) and temporally varying (i.e., weather and vegetation greenness) variables as predictors, and evaluated which predictors and aggregation periods were associated with optimal model performance.

To determine the optimal modeling framework, we tested different combinations of predictor variable sets that included all of the stationary variables, as well as temporal variables divided into different growing season stages including: (1) only the early season (referred to as "early"), (2) the early and mid season (referred to as "early-mid"), and (3) the early, mid, and late season (referred to as "early-late").

To estimate the dates corresponding to the different growing season stages, we used Covington as our target cultivar as it accounts for 85-90% of sweetpotato production in NC (Yencho et al., 2008). Togari (1950) defined early, middle and late thickening stages occurring up to 25 days after transplanting (DAT), from 25 to 60 DAT, and from 60 DAT to harvest, respectively. Similarly, Villordon et al. (2009b) found that visible storage root initiation occurs around 26 DAT. Additionally, Covington has a maturity time of 90 to 120 days (Yencho et al., 2008), and most sweetpotato slips in NC are transplanted from early May through late June (Meyers et al., 2014) and harvested from late August through early November (NC State Extension, 2017). Thus, June 1st could be considered as the average planting date across NC counties. Accordingly, in this study, the early season was defined as spanning June 1st to July 2nd (0 to 32 DAT), the mid-season as July 3rd to August 3rd (33 to 65 DAT), and the late season as August 4th to September 4th (66 to 96 DAT).

The early and early-mid models are those that could be used for in-season forecasting; i.e., the early season model could be run as early as July 2nd (before the first day of the mid-season) and the early-mid season model could be run as early as August 3rd. The late season model could not be used for in-season forecasting, since it could only be run on September 4th, which is too close to harvest so as to provide advance yield predictions. However, the early-late season model was included in this assessment for comparison purposes, particularly since we assume that late-season values (e.g., of NDVI) are more predictive of yields at harvest given that they capture conditions observed immediately before harvest.

Additionally, we tested two different composite periods for the temporal variables: 16-days and 32-days; these time periods were determined based on the temporal resolution of the satellite remote sensing data used to estimate vegetation greenness (described below). Thus, in total, there were 6 different predictor variable sets that were screened (3 based on the season times and 2 based on temporal variables).

2.3. Datasets and preprocessing

To create a generalizable model framework, we considered only predictors for which publicly available spatial datasets were available over our study period of 2008 to 2022. Annual county-scale sweetpotato yields reported by the USDA Survey (USDA National Agricultural Statistics Service, 2022b) was the response variable, and predictors were averaged over each county. We also compared the total harvested sweetpotato area reported at the county- and state-scale by the USDA Survey (annual), Census (every 5 years), and CDL (annual) to assess agreement between the datasets and identify potential bias. Because the USDA Census only occurs every 5 years, we could only consider Census data from 2012 and 2017. 2008 was chosen as the initial study year because sweetpotato was not included in the CDL prior to 2008.

As candidate predictor variables (Table 1), we considered topography and soil characteristics, which were temporally stationary, and precipitation, maximum and minimum temperature, GDD and NDVI as temporally variant predictors.

Table 1: Environmental variables used as candidate predictors in the machine learning models for sweetpotato yield forecast

Type	Variable	Source	Resolution
Stationary	Elevation (m.)	STRM V4 - CGIAR[1]	90 m.
	Slope		
	Aspect		
	Sand (%)	SoildGrids 2.0 - ISRIC[2]	250 m.
	Clay (%)		
	pH		
	Cation Exchange Capacity		
	Bulk density		
	Nitrogen		
	Soil Organic Carbon		
Temporal	Precipitation (mm.)	PRISM - Climate Group[3]	4638.3 m.
	Maximum temperature (°C)		
	Minimum temperature (°C)		
	NDVI	Landsat 5, 7, 8[4]/ Sentinel-2 - NASA[5]	30 m.
Target	Yield (t/ha)	USDA Statistics[6]	county

[1] Jarvis et al. (2008), [2] Poggio et al. (2021), [3] PRISM Climate Group, Oregon State University (2022), [4] https://landsat.gsfc.nasa.gov/, [5] https://sentinel.esa.int/web/sentinel/missions/sentinel-2, [6] https://quickstats.nass.usda.gov/

2.3.1. Stationary predictor variables: Topography and soil

Topography and soil characteristics were assumed to be stationary over the study period. For topography variables, digital elevation data (DEM) from the Shuttle Radar Topographic Mission (SRTM), produced by NASA and improved by the Consortium for Spatial Information (CGIAR-CSI) (Jarvis et al., 2008), was used to get elevation (m.), and derived to slope (°) and aspect (°) using the ee.Terrain.slope and ee.Terrain.aspect functions, respectively, from Google Earth Engine (GEE). Soil characteristics such as sand (%), clay (%), pH (phh2o), cation exchange capacity (cec, $cmol(c).kg^{-1}$), bulk density (bdod, $kg.dm^{-3}$), nitrogen ($g.kg^{-1}$), and soil organic carbon (soc,$g.kg^{-1}$)

at 5 - 15 cm depth from SoilGrids (Poggio et al., 2021), a global gridded soil information database that accounts for multiple soil characteristics at different depth ranges, were also included as predictor variables.

2.3.2. Temporal predictor variables: Weather and vegetation greenness

When summarizing the variables, we used a naming convention of the variable's abbreviation and composite start date; for instance, tmin_06-01 corresponds to the mean daily minimum temperature for the composite starting on June 1st until either June 16th or July 2nd, depending on the time aggregation or composite (i.e., 16- or 32-days).

In-season weather was accounted for using daily precipitation (ppt, mm.), and maximum (tmax, °C) and minimum temperature (tmin, °C) from the PRISM (Parameter-elevation Regressions on Independent Slopes Model) Daily Spatial Climate Dataset AN81d (PRISM Climate Group, Oregon State University, 2022), developed by the PRISM Climate Group at Oregon State University. In addition, GDD (i.e., the heat accumulation that contributes to crop growth and development) was calculated according to Equation 1 (Dufault, 1997). This equation was used based on Villordon et al. (2009a), who found that it was the more accurate GDD formulation for predicting sweetpotato yields.

$$GDD = \begin{cases} 0, & \text{if } Tmin < B \\ C - B, & \text{if } Tmin \geq B \text{ and } Tmax > C \\ Tmax - B, & \text{if } Tmin \geq B \text{ and } Tmax \leq C \end{cases} \tag{1}$$

In Equation 1, $Tmax$ and $Tmin$ are the daily maximum and minimum temperatures, respectively. B is the base temperature for total biomass production, defined as 16.9 °, and C is the ceiling temperature, defined as 29.2 ° (Gajanayake et al., 2014). Daily precipitation and GDD were summed to create totals, whereas daily maximum and minimum temperatures were averaged for every composite period.

NDVI was included as a proxy of crop health and growth, as well as a measure of unmonitored management practices (e.g., agrochemical application). Although sweetpotato grows underground, studies on other roots and tubers (including sweetpotato) report a correlation between storage root biomass and canopy growth and development, especially from root establishment to maximum canopy expansion (Tedesco et al., 2021). NDVI is the most widely-used metric for quantifying the health and density of vegetation, and it is calculated from red and near-infrared light surface reflectance (Equation 2). NDVI ranges from -1 to 1, where 1 corresponds to more dense and healthy vegetation, positive values close to 0 correspond to no vegetation (i.e., bare soil or urban areas), and negative values correspond to the presence of water.

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \tag{2}$$

NDVI was calculated from different satellites (Landsat series and Sentinel-2) and sensors to maximize data availability (Table 2). The 16- and 32-day aggregation periods for the temporal input variables were defined because of the Landsat revisit time. Building an NDVI time series using different sensors with different spectral ranges requires a harmonization process of either the surface reflectances or NDVI (Chastain et al., 2019; Villaescusa-Nadal et al., 2019; Zhang et al., 2018; Roy et al., 2016), or an equivalent atmospheric correction that allows for the intercomparison of surface reflectances across sensors (Yin et al., 2019). Thus, even though Sentinel-2 top-of-atmosphere (TOA) provided a longer period of available images (from 2015), the harmonization process of the atmospherically-corrected surface reflectance (bottom-of-atmosphere, BOA) was preferred due to its use requiring considerably less processing time.

Table 2: Sensors characteristics used for calculating NDVI time series. *Note that spatial resolution is for Red and NIR bands only.

Satellite and sensor	Spatial resolution* (m.)	Revisit time (days)	Red band, band range (nm)	NIR band, band range (nm)	Years used
Sentinel-2 MSI	10	5	B4 (650 - 680)	B8 (785-900)	2019 - 2022
Landsat 8 OLI	30	16	SR_B4 (640 - 670)	SR_B5 (850 - 880)	2013 - 2018
Landsat 7 ETM+	30	16	SR_B3 (630 - 690)	SR_B4 (770 - 900)	2012
Landsat 5 TM	30	16	SR_B3 (630 - 690)	SR_B4 (760 - 900)	2008 - 2011

Table 2 summarizes the Landsat (OLI, ETM+, TM) and Sentinel-2 (MSI) level 2 products used for the study time frame, as well as the spectral, spatial and temporal resolutions considered when calculating the NDVI time series. First, preprocessing included the conversion of Digital Numbers (DNs) to surface reflectance (scaled from 0 to 1). DNs from Landsat surface reflectance bands (collection 2) were converted to surface reflectance by multiplying pixel values by 0.0000275 and substracting 0.2 (U.S. Geological Survey, 2023). Similarly, DNs from the Harmonized Sentinel-2 MSI were multiplied by 0.0001. Afterwards, pixels with questionable data (e.g. clouds) were masked using their corresponding pixel quality band. In addition, pixels with missing values were filled using a square kernel radius of 1 pixel. This algorithm was iterated four times for Sentinel-2, Landsat TM and OLI, and 10 times for Landsat ETM+ in order to account for the black strips caused by the Scan Line Corrector (SLC) failure in 2003. Then, NDVI was calculated using Equation 1 and all images found in every time composite were aggregated by the maximum NDVI pixel value. Finally, NDVI values calculated from Landsat were harmonized to Sentinel-2 using ordinary least squares (OLS) regression coefficients shown in Table 3. The 16-day composite starting on August 4th, 2011, (ndvi_08-04) was removed due to suspiciously low NDVI values, suspected to be due to cloud cover.

Table 3: Ordinary least squares (OLS) regression coefficients to harmonized NDVI values calculated from surface reflectance images from Landsat and Sentinel-2.

Sensor	Intercept	Slope	Reference
TM / ETM+ to OLI	0.0235	0.9723	Roy et al. (2016)
OLI to MSI	0.0016	1.0016	Zhang et al. (2018)

2.3.3. Preprocessing

Yearly stationary and temporal input variables were preprocessed in the Google Earth Engine (GEE) Code Editor, an open-source cloud computing platform designed to access and analyze geospatial data. Temporal input variables followed the initial preprocessing as described in Section 2.3.2. Afterwards, all stationary and temporal potential predictors (described above) were masked such that only pixels overlapping with sweetpotato harvested areas identified by CDL were considered in the analysis, and values were then spatially averaged per county. Finally, predictors and target variable yield datasets were merged, resulting in a total of 95 records given that not all 17 counties reported yields for sweetpotatoes in the 14 years of study.

2.4. Machine learning models

ML models can handle a large number of predictors. However, including potential predictors that are redundant or not important can cause overfitting or a decrease in model performance (Khan et al., 2020). Among the various feature selection approaches that exist, the Boruta algorithm (Kursa et al., 2010), a method based on the Random Forest algorithm for identifying all relevant variables, was chosen because of its effectiveness working with remotely sensed agricultural data (Fei et al., 2022; Keskin et al., 2019). While the Boruta method is based on the Random Forest algorithm, its feature selection results are broadly applicable for use with other machine learning algorithms. In the Boruta method, the variable importance of a feature is measured by calculating the average loss of its accuracy divided by the standard deviation of all losses (Z-score) with reference to that of the shadow attributes (a randomized copy of the system variables) (Kursa et al., 2010; Kursa and Rudnicki, 2010). The Boruta method assigned predictors as relevant, tentative, and non-relevant. The tentative attributes were reanalyzed and forced to be classified as relevant or not relevant.

Feature scaling is also a key transformation in an ML pipeline since predictor variables are in different units and scales, and can cause discrepancies affecting model performance and variable importance scores. There are several scaling methods like minimum-maximum normalization and standardization; however, they can impact model performance differently (Ahsan et al., 2021). Even when tree-based algorithms do not need feature scaling because model performance is not affected, neglecting to scale predictors can affect variable importance measures (Strobl et al., 2007; Balabaeva and Kovalchuk, 2019). Therefore, the most common and successfully scaling methods - minimum-maximum normalization, standardization, and standardization combined with the YeoJohnson method - were evaluated, and the one that produced the best model performance was chosen for each model.

For the ML implementation approach, four of the most commonly used ML models for yield forecasting reported in literature (Cao et al., 2021; Van Klompenburg et al., 2020) - RFR, ANN, SVM, and XGB - were trained, tuned, tested and compared. After feature selection, the dataset was partitioned into 5-folds based on the response variable, and every model was trained and tested five times, using four folds for training and one fold for testing. Then, hyperparameter tuning employed a 10-fold cross-validation technique repeated 3 times, which is especially useful and robust for small datasets. Each model hyperparameters were manually tuned for each ML algorithm. The overall model performance was determined by computing the average metrics over the 5-fold.

Finally, a model based on all data (without partitioning) and the best-performing algorithm was trained, and the most important predictor variables affecting sweetpotato yield predictions were analyzed.

Model training, tuning, and testing were implemented in R (R Core Team, 2023) and RStudio (RStudio Team, 2021) using the Boruta (Kursa and Rudnicki, 2010), caret (Kuhn and Max, 2008), randomForest (Liaw and Wiener, 2002), nnet (Venables and Ripley, 2002), and xgboost (Chen and Guestrin, 2016) packages.

2.5. Model Evaluation Metrics

The performances of the ML models were evaluated using the root mean squared error (RMSE, Equation 3) and R squared score (R^2, Equation 4). The RMSE quantifies the difference between predicted values and actual values in the same target units, and the R^2 represents the proportion of variance explained by the model (Chicco et al., 2021). RMSE was indicated as the optimizer metric during the training process.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

In Equations 3 and 4, n is the number of data points, y_i is the observed value for the $i - th$ data point, $\hat{y}_i$ is the predicted value for the i data point, and $\bar{y}$ is the mean of the observed values y_i.

3. Results

3.1. Consistency between USDA CDL, Survey and Census data

Prior to training and testing forecast models, the harvested areas of sweetpotato were evaluated across the USDA CDL, Survey, and Census. Since the CDL was used for data preprocessing, but the

Survey yield data were used as training data for the response variable, we wanted to ensure there was reasonable consistency between the datasets with regards to their shared variable, harvested area. The comparison of annual sweetpotato harvested areas estimated by CDL with respect to the USDA Survey data (Figure 3) showed a high agreement between 2008 to 2016, with 26% more harvested areas reported by the Survey than the CDL on average. Conversely, from 2017 to 2022, CDL had a very low agreement with NASS Survey data between 2017 to 2019, reporting on average 103% more harvested area than the Survey data, and a very good agreement between 2020 to 2022, with an average 13% increase in area reported by CDL than the Survey. In addition, when comparing with USDA Census data, both CDL and survey data estimated 26% less harvested areas in 2012; and 1.7% more and 43% less harvested areas, respectively, in 2017.

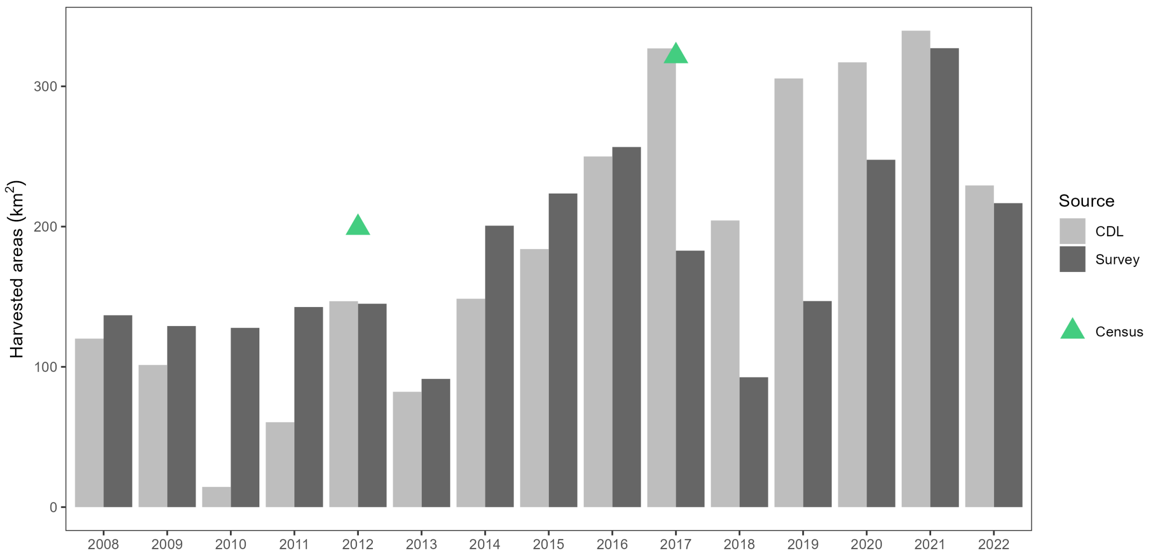


Figure 3: Sweetpotato harvested areas reported by Crop Data Layer (CDL) (USDA National Agricultural Statistics Service, 2022a), Quick Stats and the Census of Agriculture (USDA National Agricultural Statistics Service, 2022b) during the time frame of study

The CDL and Survey data were compared with Census data at the county scale, and trend lines, which intersected at zero, were fitted, resulting in line slopes of 0.93 (CDL vs. Census) and 0.58 (Survey vs. Census) in 2012, and 0.76 (CDL vs. Census) and 0.76 (CDL vs. Census) in 2017 (Table 4). Top producing counties were similar for all data sources in both years; therefore, CDL and Survey data had high correlation ($R^2 \geq 0.88$) when compared with Census data. The agreement between CDL and Survey data provided support to the utilization of CDL data for preprocessing predictors and Survey data as the model's true target yield.

3.2. Feature selection

The Boruta feature selection applied to the six predictor sets showed the same top three most important variables - elevation, nitrogen and cec - and somewhat less important variables, soc and

Table 4: Comparison of the sweetpotato harvested areas per county reported by Census, Survey and Crop Data Layers (CDL on Census of Agriculture years (2012 and 2017). Descending ordered from Census data. Showing only harvested areas $> 1km^2$

2012				2017			
County	Census	Survey	CDL	County	Census	Survey	CDL
Johnston	43.05	37.64	20.63	Nash	62.74	39.13	32.18
Nash	41.59	37.84	24.33	Johnston	60.17	43.71	40.74
Sampson	37.35	36.22	25.64	Sampson	55.50		54.32
Wilson	27.73	28.13	16.13	Wilson	45.41	39.26	36.73
Wayne	11.04		7.08	Edgecombe	17.13	34.28	26.90
Columbus	10.23		3.35	Wayne	15.37		27.64
Edgecombe	6.35		7.88	Duplin	14.36	13.76	12.93
Duplin	6.22		7.51	Harnett	12.57		9.43
Harnett	4.77	5.18	3.08	Wake	11.25		2.83
Pitt	4.74		4.91	Columbus	9.64		5.18
Robeson	2.57		2.14	Lenoir	6.34	11.13	11.31
Wake	2.48		1.45	Robeson	5.09		3.02
Chowan	1.09			Pitt	4.41		16.96
Cumberland			3.38	Lee	0.98		
Greene			9.63	Moore	0.68		
Lenoir			5.44	Martin		1.58	1.55
				Bertie			3.93
				Halifax			3.53
				Cumberland			3.29
				Scotland			1.33
				Bladen			1.24
				Greene			23.43
Total	199.19	145.00	142.58	Total	321.63	182.84	318.47
Slope		0.93	0.58	Slope		0.76	0.76
R^2		1.00	0.96	R^2		0.91	0.88

clay (Figure 4). Consistently, all predictor sets had similar variables following the top three most important variables, which were temporally variant variables including NDVI and GDD at specific time points in the mid-season. However, the specific time points corresponding to important predictors varied depending on the data configuration. Figure 4 illustrates variable importance along with the final classification of variables (unimportant, important) for early-mid predictor sets (16- and 32-days composites).

3.3. Model performance and selection of final model

The 5-fold average metrics (Table 5) showed that RFR consistently outperformed other ML algorithms. The best model was selected based on the R^2 and RMSE from the testing data, prioritizing the early and early-mid season models over late season models given their ability to provide in-season forecasts well ahead of harvest. Thus, the early-mid season with 16-day aggregation had the best testing (RMSE = 3.53 $t.ha^{-1}$, R^2 = 0.44) performance, which was exactly the same as the early-late model.

XGB performed very similar or even slightly better than RFR during both training and testing, but only when temporal variables included data through the late season. However, when temporal variables included data only up to the early or mid season, XGB performance decreased and fell below

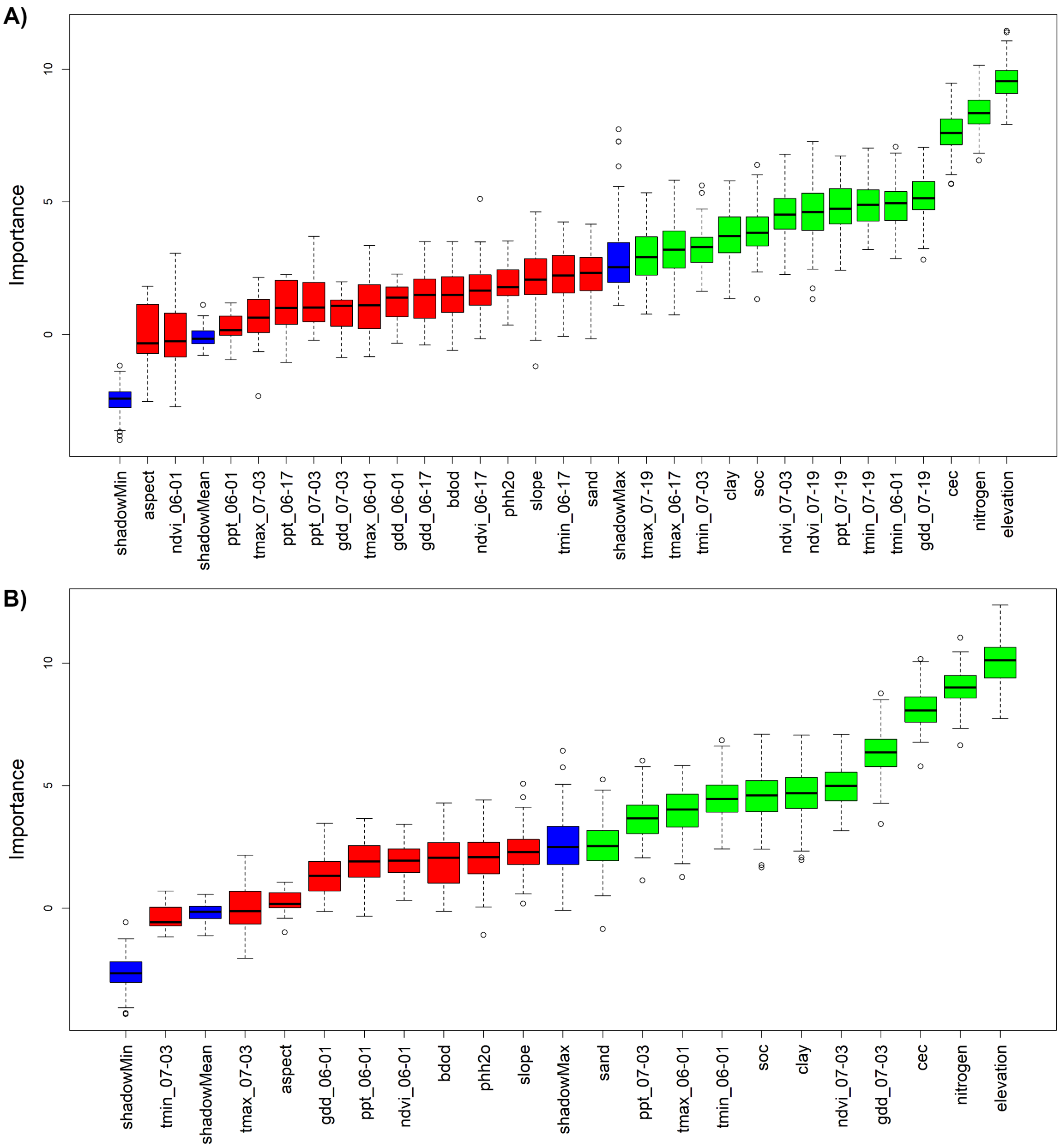


Figure 4: Importance of all potential predictors, determined by the Boruta Feature selection method for stationary and temporal variables until Mid season (June - July) with A) 16-days and B) 32-days time composites. Where, the most important variables are in green, the least important variables are in red, and the threshold variables are in blue. See Supplementary Materials for Early (June) and Late (June - August) seasons.

that of RFR.

Table 5: Model performance, Root Mean Square Error (RMSE, $t.ha^{-1}$) and R^2, of county-level sweetpotato yield forecast for different ML algorithms, input time composites and season. ML models Random Forest (RFR), Artifitial Neural Networks (ANN), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGB) were applied using stable and temporal variables until late (June - August) and Mid (June - July) season, with 16-days and 32-days time composites.

Composite	Season	Model	Training		Testing	
			RMSE	R^2	RMSE	R^2
16-days	early -late	RFR	1.58	0.89	3.52	0.44
		ANN	2.47	0.72	3.92	0.29
		SVM	2.90	0.62	4.02	0.27
		XGB	1.25	0.92	3.53	0.43
	early - mid	RFR	1.58	0.89	3.53	0.44
		ANN	2.67	0.68	3.94	0.30
		SVM	2.85	0.63	4.11	0.23
		XGB	1.10	0.93	3.77	0.36
	early	RFR	1.70	0.87	3.77	0.36
		ANN	3.49	0.46	4.33	0.15
		SVM	3.02	0.58	4.30	0.14
		XGB	1.98	0.78	4.04	0.26
32-days	early -late	RFR	1.61	0.88	3.53	0.44
		ANN	2.77	0.65	3.92	0.31
		SVM	2.91	0.61	4.05	0.25
		XGB	1.23	0.89	3.46	0.46
	early - mid	RFR	1.68	0.87	3.67	0.39
		ANN	3.27	0.52	4.18	0.21
		SVM	3.05	0.58	4.44	0.10
		XGB	2.18	0.79	3.49	0.44
	early	RFR	1.66	0.88	3.65	0.4
		ANN	3.40	0.49	4.27	0.17
		SVM	3.32	0.51	4.40	0.12
		XGB	1.87	0.8	3.71	0.38

Considering both performance and ability to be used as an in-season forecasting model, the RFR algorithm built with early-mid season predictors aggregated at 16-days was selected as the "best" model and further analyzed. Due to the small dataset, a final RFR model was trained without data partitioning using out-of-bag (OOB) cross validation to maximize the amount of data available for model training. The final model's OOB error was RMSE = 3.64 $t.ha^{-1}$ with a R^2 = 0.41. The observed versus OOB prediction plot (Figure 5) depicted some decrease in yield, which we noted varied as a function of elevation.

3.4. Predictor variable importance

The most important predictors in the final model (Figure 6) included the stationary variables of elevation, nitrogen, and cec, since without accounting for them, the prediction error (OBB MSE) would increase by 14, 11 and 10 $t.ha^{-1}$, respectively. Temporal variables ndvi_07-03, ppt_07-19 and tmin_06-01 were also deemed important, as permuting each of them resulted in an error of 9.7, 8.7 and 7.5 $t.ha^{-1}$, respectively. The variable ndvi_07-03 represented the vegetation greenness in sweetpotato

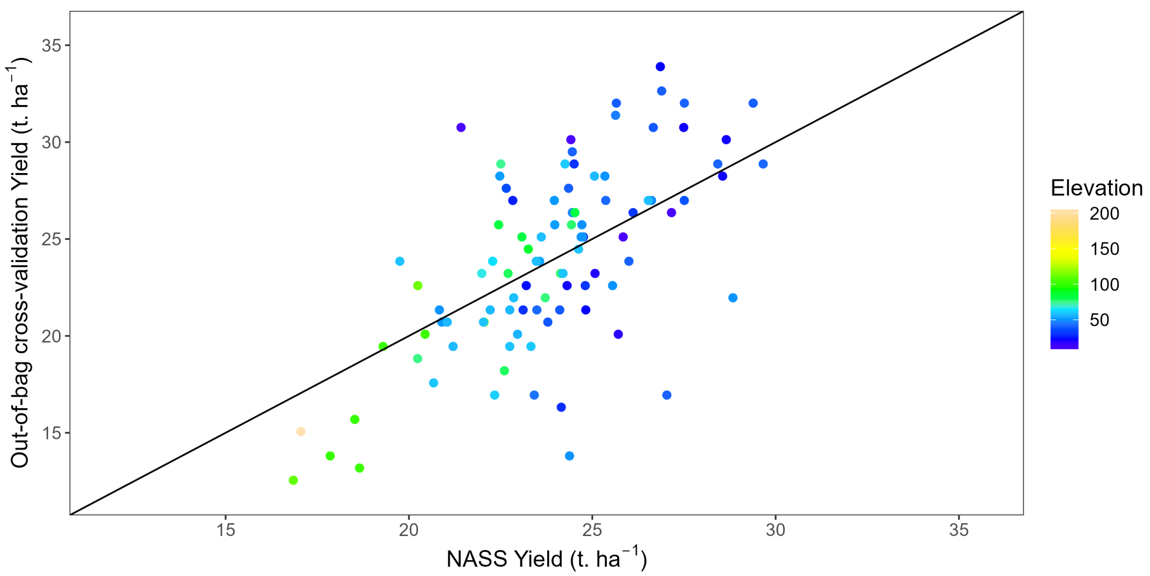


Figure 5: NASS county yields (USDA National Agricultural Statistics Service, 2022b) versus out-of-bag predictions from the final model. Dots are colored by the most important variable, elevation (m.)

pixels within a county during the first month of the growing season, which is the critical stage for storage root initiation. The variable ppt_07-19 was the amount of precipitation during storage bulking, which also influences final yield (Gajanayake and Reddy, 2016). And, tmin_06-01 included the minimum temperature around planting and establishment, also a critical stage.

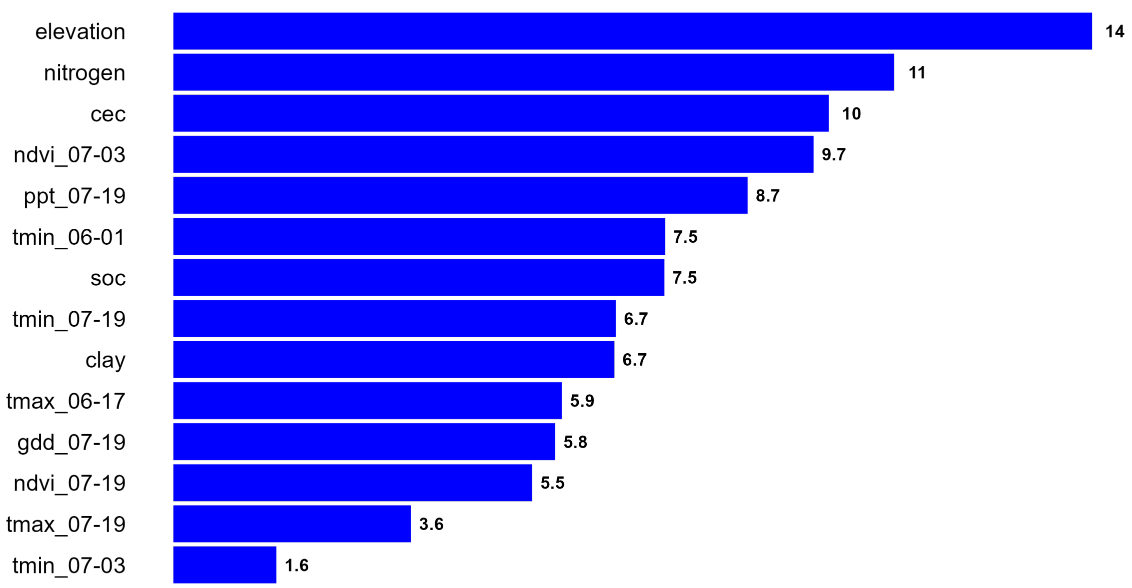


Figure 6: Importance of predictors determined by the Random Forest model (RFR) for sweetpotato yield forecasting at county-level. The RFR model was built with stationary and early to mid season predictors and 16-day composites for temporally variant predictors. Importance is defined as the increase in the MSE prediction when the variables is permuted (e.g. 14 for elevation)

Figure 7 depicts the variability of the variables included in the final RFR model and the target variable. While the response variable, yield, ranged from 12.55 to 33.89 $t.ha^{-1}$ without outliers, some predictors (e.g., elevation, soc and gdd_07-19) had low spatial variation, with some outliers.

Unsurprisingly, since high production areas for sweetpotato in North Carolina are primarily situated in the Coastal Plains, with elevations gradually increasing towards the northeast, county elevations ranged from 9 to 205 meters above sea level (m.a.s.l.), with approximately 50% of the data falling within the range of 39 to 60 m.a.s.l. In contrast, nitrogen and cec at 5-15 cm depth ranged from 0.78 to 1.70 $g.kg^{-1}$ and from 6.4 to 16.3 $cmol_c.kg^{-1}$, respectively, had a more balanced distribution with only a few superior outliers. The cec values were distributed across the typical values for fine sandy loam and loam soil, with low to medium organic matter and water holding capacity. NDVI values after storage root initiation (ndvi_07-03) ranged from 0.19 to 0.81; however, 50% of data ranged only from 0.39 to 0.53, which means that vegetation coverage was about the same. Similarly, total precipitation during the period of greatest storage root bulking (ppt_07-19), ranged from 4.58 to 197.19 mm., showing the high variability of rain even within a relatively small productive region. Minimum temperature just after planting mostly varied from 18.76 to 19.85 °C (tmin_06-01), with a few minimum outliers down to 14.82 °C.

4. Discussion

When screening candidate predictor variables, we found that the stationary variables of elevation, nitrogen, and cec consistently had the highest importance, and soc and clay to a lesser extent (Figure 4. In the final model built with RFR, the stationary variables had the greatest effect on sweetpotato yield predictions at the county scale. Elevation was the most important predictor variable; however, it should be interpreted as an indicator of the geographic location in NC, since sweetpotatoes are grown in a region with flat terrain. Similarly, when considering the factors that influence soil formation across the state, elevation emerges as the most influential element, significantly shaping the definition of soil units and, consequently, their characteristics (Lee, 1955). Thus, elevation is most likely acting as a proxy for other geospatial covariates such as soil quality and climate patterns, which spatially and temporally determine the sweetpotato growth. As a result, these factors make certain regions more suitable for sweetpotato cultivation, leading to the development of more advanced management practices, and consequently, achieving high yields. With regards to the importance of nitrogen, since it represents the total nitrogen in the soil rather than the nitrogen fertilizer applied or available to the plants, it is likely a proxy for soil drainage quality. Well-drained soils often require more nitrogen application due to leaching. These soils typically are sandy textures, which are preferred by sweetpotatoes due to the ample pore spaces that facilitate storage root growth. Finally, cec's importance reflects its role in indicating soil type and health, as soils with higher cec are better at retaining essential nutrients (Kaiser et al., 2008), directly correlating with the optimal growth and yield of sweetpotatoes in NC's diverse agricultural systems. The predominant importance of elevation and soil properties (stationary variables) over weather or vegetation indexes (temporal variables) is

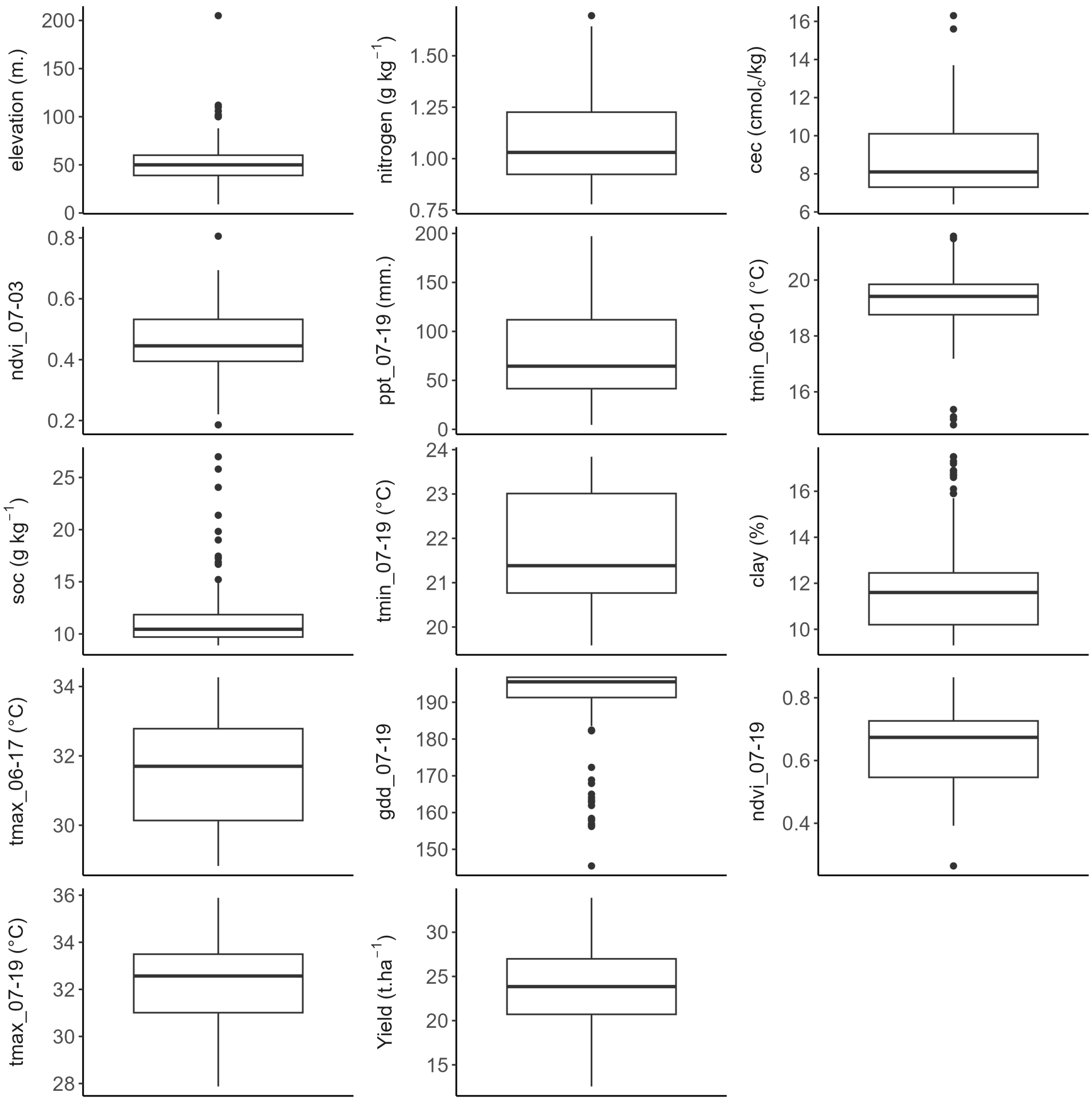


Figure 7: Boxplot showing the variability of all-relevant variables selected after Boruta feature selection

supported by previous studies such as Cao et al. (2021), where elevation had significant correlations between climate factors when modeling wheat yield for thirteen China's provinces.

In contrast to stationary predictors, temporal variables were less consistently important, particularly when data across multiple growing season stages (i.e., early-mid, early-late) were considered. In the final RFR model, which only considered measurements of temporally-varying predictors from the early and mid growing season, temporal input variables ndvi_07-03, ppt_07-19, and tmin_06-01 were the most important temporally variant predictors. The importance of ndvi_07-03, the 16-day NDVI composite starting on July 03, the 4th most important variable, indicates that canopy growth is an important predictor of end-of-season yields. The importance of NDVI in the early and mid-growing season as a yield predictor is corroborated by prior research demonstrating that early and mid-canopy growth is correlated with root development (Tedesco et al., 2021). Variable ppt_07-19, the 5th most important variable, represented the total precipitation in the mid-season, which is a critical time for storage root bulking and the end-season yield (Gajanayake and Reddy, 2016), with some variability across regions in NC (Zarzar and Dyer, 2019). Finally, tmin_06-01, the 6th most important variable, is the mean minimum temperature in the early growing season near transplanting. Because average temperatures need to be greater than 16.8 ° for successful sweetpotato transplanting and root establishment (Gajanayake et al., 2014), the inclusion of this predictor in the final model captures a known mechanism driving sweetpotato growth and yields. Overall, the predictor variables and the associated timing of the predictors included in the final model correspond to established environmental relationships known to affect sweetpotato productivity.

Though the RFR algorithm with 16-day composite predictors spanning the early and mid-growing season was selected as the best and final model, other ML algorithms and predictor composite periods were considered. RFR outperformed the other ML algorithms (ANN, SVM and XGB) and its evaluation metrics were the most consistent across the six considered predictor sets, especially when both mid and late-season temporal variables were included as predictors. Models with 16- and 32-day composite and only early season data had slightly lower performances. This suggests that having numerous temporal variable predictors was not necessarily advantageous; instead, the composited data may have introduced noise into the model, diminishing its robustness. The final model's performance was moderate (testing: R^2 = 0.44, RMSE = 3.53 $t.ha^{-1}$), and considered acceptable for forecasting given that the models predict how sweetpotato yield will vary as a function of environmental conditions alone. To improve model performance, predictors capturing other drivers that affect actual yield (e.g., genotype characteristics and management practices) should be included. Future work could build upon the regional models presented here to provide more tailored forecast products for individual farms.

Interestingly, models that included predictors from the late growing season did not outperform the models that only considered temporally varying predictors from the early and mid-season. The inclusion of late-season predictors was expected to result in more accurate forecasts since conditions late in the growing season (e.g., NDVI near harvest) were expected to more closely correlate with final yields. Yet, the testing R^2 for the RFR model with 16-day composites and early-late season predictors was 0.44, and the equivalent model with early-mid season predictors was also 0.44, indicating the late season values did not improve model performance. However, the testing R^2 for the RFR model with 16-day composites that only included early season predictors was 0.36. These results indicate that environmental conditions spanning the early and mid-season are predictive of yields at harvest, and that information from later in the season is not necessary to boost performance. Operationally, these findings demonstrate that the best time to run this forecast model is at the end of the mid-season, approximately between four to eight weeks before harvest. However, depending on end-user interests, the RFR model with 16-day composites using only early season values for temporal predictors (i.e., the model with a testing R^2 of 0.36) may be more desirable despite its poorer performance, as it could be run as early as eight to twelve weeks before harvest.

While our results demonstrate an in-season yield forecasting model for sweetpotato performs reasonably well and can provide actionable information, there are opportunities to further expand the frameworks tested here. For example, Zhou et al. (2022) found that solar-induced chlorophyll fluorescence (SIF) had better predictability for yield than traditional vegetation indices, so the inclusion of SIF could potentially improve model performance. Similarly, previous studies reported superior predictive power from Land Surface Temperature (LST) over air temperature given it provides information on canopy temperature, which is related to water and heat stresses (Kang et al., 2020; Siebert, Stefan and Ewert, Frank and Rezaei, Ehsan Eyshi and Kage, Henning and Graß, Rikard, 2014; Pede et al., 2019). Although deep learning (DL) algorithms, such as Long Short-Term Memory (Van Klompenburg et al., 2020), have shown promise in yield forecasting at county scale, they typically require larger datasets. Furthermore, a prior study comparing DL and traditional machine learning (ML) algorithms found that DL algorithms did not demonstrate superior performance over ML models like Random Forest or Extreme Gradient Boosting at the county scale (Kang et al., 2020; Cao et al., 2021).

The most significant limitation of this study stemmed from the small sample size, which was a result of the restricted availability of yield data for sweetpotatoes, thereby constraining the ML process and its overall robustness. Additionally, at the county level, spatial error and uncertainties were inevitably introduced to the model during data preprocessing, particularly when compositing temporal data and matching coarse spatial data of varying resolution. Moreover, as demonstrated by the comparison of sweetpotato harvested areas from USDA CDL, Survey, and Census data (Figure 3 and Table 4), there are uncertainties even in the locations of farms, and the reliance on farmerreported yields results in the sweetpotato yield data being affected by survey participation rates and respondent honesty. Additionally, while satellite image inputs and classification methods have led to improved CDL accuracy over time, the CDL's accuracy has only been tested for select crops and regions; prior research shows that CDL performs best for major crops (producer's accuracies of 71.5%), aggregated categories, and within major cropping regions such as the Corn Belt, Central Plains, and Mississippi Delta (Lark et al., 2021). Regardless, CDL demonstrated good county-scale agreement with Census and Survey data (Table 4), and was considered a good source for identifying sweetpotato fields since it is able to estimate field locations that may be hidden from NASS survey data in an effort to protect grower identity. NASS safeguards individual farm privacy by excluding farms from reporting if they have fewer than 100 planted acres, and about 50% of farms in North Carolina are between 1 to 49 acres (0.4 - 19.8 *ha*) (U.S. Department of Agriculture, 2022).

5. Conclusions

This study analyzed four ML algorithms for predicting sweetpotato yield using stationary and temporal environmental variables as potential predictors. The six predictor sets, which varied in the amount of in-season data they considered as well as the aggregation period of temporal variables (16- vs 32-day), provided key information about important variables and models' performance. In particular, elevation, which is an indicator of geographic location, had the highest importance. The RFR model consistently outperformed the other ML algorithms. We determined that the best model configuration for temporal variables used early and mid-season data with 16-day composited temporal variables. Using late-season data did not improve model performance. Among the various input variables considered, the stationary ones (elevation, nitrogen and cec), followed by NDVI and precipitation after storage root initiation and bulking (July), and minimum temperature around planting (June), were the most predictive of sweetpotato yield at the county scale.

Publicly available in-season remote sensing data, coupled with machine learning models, can predict sweetpotato yields reasonably well before the season's end. This approach aims to aid growers in enhancing their harvest management, optimizing marketable yield, planning storage, refining sales and marketing strategies, and even plan for the next year's planting. Furthermore, it provides valuable insights to decision-makers, facilitating more accurate estimates of crop insurance payments, revenue support programs, and collaborative planning with local extension agents and agribusinesses.

Acknowledgements

This work was supported by USDA NIFA project 2019-67021-29936, grant (1404) 2020–1160 through North Carolina State University's Plant Sciences Initiative (NC PSI) Game-Changing Research Incentive Program, and USDA NIFA Hatch project 7003506.

References

Abbas, F., Afzaal, H., Farooque, A.A., Tang, S., 2020. Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy 10, 1046.

Ahsan, M.M., Mahmud, M.P., Saha, P.K., Gupta, K.D., Siddique, Z., 2021. Effect of data scaling methods on machine learning algorithms and model performance. Technologies 9, 52.

Balabaeva, K., Kovalchuk, S., 2019. Comparison of temporal and non-temporal features effect on machine learning models quality and interpretability for chronic heart failure patients. Procedia Computer Science 156, 87–96.

Cao, J., Zhang, Z., Luo, Y., Zhang, L., Zhang, J., Li, Z., Tao, F., 2021. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. European Journal of Agronomy 123, 126204.

Chastain, R., Housman, I., Goldstein, J., Finco, M., Tenneson, K., 2019. Empirical cross sensor comparison of sentinel-2a and 2b msi, landsat-8 oli, and landsat-7 etm+ top of atmosphere spectral characteristics over the conterminous united states. Remote sensing of environment 221, 274–285.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 785–794. URL: http://doi.acm.org/10.1145/2939672.2939785, doi:10.1145/2939672.2939785.

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. PeerJ Computer Science 7, e623.

Dufault, R.J., 1997. Determining Heat Unit Requirements for Broccoli Harvest in Coastal South Carolina. Journal of the American Society for Horticultural Science 122, 169–174. doi:10.21273/JASHS.122.2.169. publisher: American Society for Horticultural Science Section: Journal of the American Society for Horticultural Science.

Fei, S., Li, L., Han, Z., Chen, Z., Xiao, Y., 2022. Combining novel feature selection strategy and hyperspectral vegetation indices to predict crop yield. Plant Methods 18, 1–13.

Gajanayake, B., Reddy, K.R., 2016. Sweetpotato responses to mid-and late-season soil moisture deficits. Crop Science 56, 1865–1877.

Gajanayake, B., Reddy, K.R., Shankle, M.W., Arancibia, R.A., Villordon, A.O., 2014. Quantifying Storage Root Initiation, Growth, and Developmental Responses of Sweetpotato to Early Season Temperature. Agronomy Journal 106, 1795–1804. URL: https:

//onlinelibrary.wiley.com/doi/abs/10.2134/agronj14.0067, doi:10.2134/agronj14.0067. _eprint: https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2134/agronj14.0067.

Ghazaryan, G., Skakun, S., König, S., Rezaei, E.E., Siebert, S., Dubovyk, O., 2020. Crop yield estimation using multi-source satellite image series and deep learning, in: IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 5163–5166.

Gómez, D., Salvador, P., Sanz, J., Casanova, J.L., 2019. Potato yield prediction using machine learning techniques and sentinel 2 data. Remote Sensing 11, 1745.

Griffith, G.E., Omernik, J.M., Comstock, J., Schafale, M., McNab, W., Lenat, D., MacPherson, T., 2002. Ecoregions of North Carolina. Western Ecology Division, National Health and Environmental Effects Research ....

Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., et al., 2008. Hole-filled srtm for the globe version 4. available from the CGIAR-CSI SRTM 90m Database (http://srtm. csi. cgiar. org) 15, 5.

Jung, J., Maeda, M., Chang, A., Bhandari, M., Ashapure, A., Landivar-Bowles, J., 2021. The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. Current Opinion in Biotechnology 70, 15–22.

Kaiser, M., Ellerbrock, R., Gerke, H., 2008. Cation exchange capacity and composition of soluble soil organic matter fractions. Soil Science Society of America Journal 72, 1278–1285.

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., Anderson, M., 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the us midwest. Environmental Research Letters 15, 064005.

Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58.

Khan, N.M., Madhav C., N., Negi, A., Thaseen, I.S., 2020. Analysis on improving the performance of machine learning models using feature selection technique, in: Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 2, Springer. pp. 69–77.

Kuhn, Max, 2008. Building predictive models in r using the caret package. Journal of Statistical Software 28, 1–26. URL: https://www.jstatsoft.org/index.php/jss/article/view/v028i05, doi:10.18637/jss.v028.i05.

Kursa, M.B., Jankowski, A., Rudnicki, W.R., 2010. Boruta–a system for feature selection. Fundamenta Informaticae 101, 271–285.

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. Journal of Statistical Software 36, 1–13. URL: https://doi.org/10.18637/jss.v036.i11.

Lark, T.J., Schelly, I.H., Gibbs, H.K., 2021. Accuracy, bias, and improvements in mapping crops and cropland across the united states using the usda cropland data layer. Remote Sensing 13, 968.

Lee, W.D., 1955. The Soils of North Carolina: Their Formation, Identification and Use. 115, North Carolina Agricultural Experiment Station.

Leng, G., Hall, J.W., 2020. Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. Environmental Research Letters 15, 044027.

Li, D., Miao, Y., Gupta, S.K., Rosen, C.J., Yuan, F., Wang, C., Wang, L., Huang, Y., 2021. Improving potato yield prediction by combining cultivar information and uav remote sensing data using machine learning. Remote Sensing 13, 3322.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2, 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

Meyers, S.L., Jennings, K.M., Monks, D.W., 2014. 'covington'sweetpotato tolerance to flumioxazin applied post-directed. Weed Technology 28, 163–167.

NC State Extension, 2017. North carolina sweet potatoes. URL: https://lee.ces.ncsu.edu/2017/12/north-carolina-sweet-potatoes/. accessed on October 02, 2023.

Paudel, D., Boogaard, H., de Wit, A., van der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S., Athanasiadis, I.N., 2022. Machine learning for regional crop yield forecasting in europe. Field Crops Research 276, 108377.

Pede, T., Mountrakis, G., Shaw, S.B., 2019. Improving Corn Yield Prediction Across the U.S. Corn Belt by Replacing Air Temperature with Daily MODIS Land Surface Temperature. Agricultural and Forest Meteorology 276, 107615.

Pérez-Pazos, J.V., Rosero, A., Martínez, R., Pérez, J., Morelo, J., Araujo, H., Burbano-Erazo, E., 2021. Influence of morpho-physiological traits on root yield in sweet potato (ipomoea batatas lam.) genotypes and its adaptation in a sub-humid environment. Scientia Horticulturae 275, 109703.

Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G., Kempen, B., Ribeiro, E., Rossiter, D., 2021. Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7, 217–240.

PRISM Climate Group, Oregon State University, 2022. https://prism.oregonstate.edu. URL: https://prism.oregonstate.edu. accessed on September 26, 2023.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Roy, D.P., Kovalskyy, V., Zhang, H., Vermote, E.F., Yan, L., Kumar, S., Egorov, A., 2016. Characterization of landsat-7 to landsat-8 reflective wavelength and normalized difference vegetation index continuity. Remote sensing of Environment 185, 57–70.

RStudio Team, 2021. RStudio: Integrated Development Environment for R. RStudio, PBC. Boston, MA. URL: http://www.rstudio.com/.

Salvador, P., Gómez, D., Sanz, J., Casanova, J.L., 2020. Estimation of potato yield using satellite data at a municipal level: A machine learning approach. ISPRS International Journal of Geo-Information 9, 343.

Siebert, Stefan and Ewert, Frank and Rezaei, Ehsan Eyshi and Kage, Henning and Graß, Rikard, 2014. Impact of Heat Stress on Crop Yield—On the Importance of Considering Canopy Temperature. Environmental Research Letters 9, 044012.

Soto-Caro, A., Luo, T., Wu, F., Guan, Z., 2022. The U.S. sweet potato market: Price response and impact of supply shocks. Horticulturae 8, 856.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8, 1–21.

Sun, C., Feng, L., Zhang, Z., Ma, Y., Crosby, T., Naber, M., Wang, Y., 2020. Prediction of end-of-season tuber yield and tuber set in potatoes using in-season uav-based hyperspectral imagery and machine learning. Sensors 20, 5293.

Tedesco, D., de Oliveira, M.F., dos Santos, A.F., Silva, E.H.C., de Souza Rolim, G., da Silva, R.P., 2021. Use of remote sensing to characterize the phenological development and to predict sweet potato yield in two growing seasons. European Journal of Agronomy 129, 126337.

Togari, Y., 1950. A study in the tuberous-root formation of sweet potato. Bull. Natl. Agric. Exp. Stn. 68, 1–96.

U.S. Department of Agriculture, 2022. North carolina annual statistical bulletin highlights. URL: https://www.nass.usda.gov/Statistics_by_State/North_Carolina/Publications/Annual_Statistical_Bulletin/AgStat/NCHighlights.pdf. accessed on October 9, 2023.

U.S. Geological Survey, 2023. How do i use a scale factor in landsat level-2 science products? https://www.usgs.gov/faqs/how-do-i-use-a-scale-factor-landsat-level-2-science-products. Accessed on September 26, 2023.

USDA National Agricultural Statistics Service, 2022a. Cropland data layer (2008 - 2022). published crop-specific data layer. URL: https://nassgeodata.gmu.edu/CropScape/. accessed on September 26, 2023.

USDA National Agricultural Statistics Service, 2022b. NASS - Quick Stats. https://data.nal.usda.gov/dataset/nass-quick-stats. Accessed on September 27, 2023.

Van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture 177, 105709.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Fourth ed., Springer, New York. URL: https://www.stats.ox.ac.uk/pub/MASS4/. iSBN 0-387-95457-0.

Villaescusa-Nadal, J.L., Franch, B., Roger, J.C., Vermote, E.F., Skakun, S., Justice, C., 2019. Spectral adjustment model's analysis and application to remote sensing data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, 961–972.

Villordon, A., Clark, C., Ferrin, D., LaBonte, D., 2009a. Using Growing Degree Days, Agrometeorological Variables, Linear Regression, and Data Mining Methods to Help Improve Prediction of Sweetpotato Harvest Date in Louisiana. HortTechnology 19, 133–144. URL: https://journals.ashs.org/horttech/view/journals/horttech/19/1/article-p133.xml, doi:10.21273/HORTSCI.19.1.133. publisher: American Society for Horticultural Science Section: HortTechnology.

Villordon, A., LaBonte, D., Firon, N., 2009b. Development of a simple thermal time method for describing the onset of morpho-anatomical features related to sweetpotato storage root formation. Scientia horticulturae 121, 374–377.

Villordon, A., Sheffield, R., Rojas, J., Chiu, Y.L., 2011. Development of simple bayesian belief and decision networks as interactive visualization tools for determining optimal in-row spacing for 'beauregard'sweetpotato. HortScience 46, 1588–1597.

Villordon, A., Solis, J., LaBonte, D., Clark, C., 2010. Development of a prototype bayesian network model representing the relationship between fresh market yield and some agroclimatic variables known to influence storage root initiation in sweetpotato. HortScience 45, 1167–1177.

Yencho, G.C., Pecota, K.V., Schultheis, J.R., VanEsbroeck, Z.P., Holmes, G.J., Little, B.E., Thornton, A.C., Truong, V.D., 2008. 'covington'sweetpotato. HortScience 43, 1911–1914.

Yin, F., Lewis, P.E., Gomez-Dans, J.L., Wu, Q., 2019. A sensor-invariant atmospheric correction method: Application to Sentinel-2/MSI and Landsat 8/OLI. Preprint .

Zarzar, C., Dyer, J., 2019. The influence of synoptic-scale air mass conditions on seasonal precipitation patterns over North Carolina. Atmosphere 10, 624.

Zhang, H.K., Roy, D.P., Yan, L., Li, Z., Huang, H., Vermote, E., Skakun, S., Roger, J.C., 2018. Characterization of sentinel-2a and landsat-8 top of atmosphere, surface, and nadir brdf adjusted reflectance and ndvi differences. Remote sensing of environment 215, 482–494.

Zhou, W., Liu, Y., Ata-Ul-Karim, S.T., Ge, Q., Li, X., Xiao, J., 2022. Integrating climate and satellite remote sensing data for predicting county-level wheat yield in china using machine learning methods. International Journal of Applied Earth Observation and Geoinformation 111, 102861.