County-Level Crop Yield Forecast with Deeping Learning and Satellite Data

Yanghui Kang

Background

As maintaining food security for the entire population becomes more and more challenging, reliable estimation of crop yield is more imperative than ever for the scientific community (Foley et al., 2011; Lobell, 2013). The timeliness, accuracy, and spatial scales of such prediction are important factors that influence decisions making about grain market and crop insurances.

Publicly available predictions of crop yield are mainly at country or state/province level and are derived based on historical statistical relationships and in-season surveys. For example, the United States Department of Agriculture (USDA) publishes yield predictions of major staple crops for the US at monthly schedules before harvest. The average forecast error for the whole country is within ±5% in September, and ±3% in October (Irwin, 2014).

Further improvement of the utility of crop yield prediction is associated with two major goals: increasing prediction accuracy and decreasing the spatial size of prediction unit, i.e. from country/state to county or field level. At county level, research strides mainly focus on the use of satellite images as input data and machine learning algorithms as estimator (Johnson, 2014; Guan et al., 2017; You et al., 2017). Remote sensing images make objective measurements of the entire earth surface at regular time steps (from 3-days to half a month). It not only senses the visible wavelength reflected from the crop canopy, but also other longer wavelengths such as infrared that contains more information about crop biochemical prosperities. Therefore, satellite images have been historically used to establish statistical relationships with crop yields. It has been shown that remote sensing derived data could explain more that 60% of the variances in county-level crop yields (Bolton & Friedl, 2013; López-Lozano et al., 2015; Guan et al., 2017).

Recent development in artificial intelligence has seen a phenomenal improvement in the performance of machine learning, especially, deep learning algorithms (LeCun et al., 2015). Therefore, some very recent studies started to explore these algorithms and the results were encouraging (Johnson, 2014; Kuwata & Shibasaki, 2015; You et al., 2017). Johnson (2014) built regression tree estimators to predict maize and soybean yields in the US, with 8-day interval Normalized Difference Vegetation Index (NDVI) data and Land Surface Temperature (LST) data derived from Moderate Resolution Imaging Spectroradiometer (MODIS). The coefficients of determination ($R^2$) reached 0.93 for both maize and soybean if data within the whole season was used, while if only early- to mid-season data was considered, $R^2$ is above 0.7, which is still much better that the traditional linear/non-linear regression methods (Bolton & Friedl, 2013). Most
recently, You et al. (2017) applied deep learning methods, i.e. Convolutional Neural Network (CNN) and Long-short Term Memory (LSTM), to predict soybean yields for the US. They used a novel approach to transform high-dimensional, multi-band time-series MODIS images into image histograms and incorporated spatio-temporal covariance of yield through a Gaussian Process. The after-season estimation yielded only a 10% error at county-level, and the within-season prediction at country level outperformed USDA predictions by 15%. These new researches on machine learning shows that satellite data contains more information about crop yield than previous believed, possibly because many hidden, deep connections between the time series of spectral data and final crop yield cannot be revealed by simple linear or non-linear regression.

While previous studies mainly focused on remote sensing data, weather and soil information has not yet been adequately explored. It is important to note that remote sensing data is a manifest observation of the crop growth status (i.e. canopy foliage amount, chlorophyll, and water content), while the intrinsic driving factors and input components of crop yield are essentially weather (solar radiation, temperature, precipitation), soil, and biotic conditions, as well as field management by farmers. Johnson (2014) originally considered precipitation but found it to have minimum correlation with yield, and thus dropped it in the final model. You et al. (2017) considered these intrinsic factors in an indirect manner, by including a spatio-temporal term as Gaussian Process on top of the networks. They showed that adding this process improved model performance by up to 27%.

**Research Objectives**

The goal of this project is two-folds:

1. Evaluate the performance of deep learning algorithms in predicting crop yield by directly incorporating weather and soil data with remote sensing observations
2. Investigate the timeliness of crop yield forecast with only early- and/or mid- season input data.

**Proposed Methodology**

**Data**

The crop yield dataset is composed of county-level after-season average maize yield from USDA for 11 states in the US from 2001 to 2016 (). The remote sensing observations include MODIS 16-day 250m Enhanced Vegetation Index (EVI) products (MOD13Q1) and 8-day 1-km Land Surface Temperature (LST) products (MOD11A2). Weather data comes from the DAYMET gridded daily weather observations including minimum and maximum temperature, precipitation amount, and shortwave radiation at 1km resolution (Thornton et al., 2017). Soil properties will be extracted from Unified North American Soil Map (UNASM), which includes soil depth, soil texture, organic carbon content, PH, cation exchange capacity, and bulk density for the topsoil layer (0 – 30cm) and the subsoil layer (30 – 100cm) at 0.25 spatial degree (roughly 30km) (Liu et al., 2014). Each of these variables will be spatially aggregated to mean and standard deviation within the
boundary of each county, while considering maize crop mask from Crop Data Layer (CDL) (Johnson & Mueller, 2010).

Deep Learning Algorithms

For this project, Recurrent Neural Network (RNN) and Multiple-Layer Perception (MLP) algorithms will be evaluated. For RNN, the sequential data, such as mean and standard deviation of remote sensing observations and weather data will be resampled to 8-day time steps. For simplicity, soil properties, albeit non-sequential in nature, will be treated as a series of data of the same value across time. For MLP, considering the dimension of feature space and relative scarcity of training yield data (~6000), the sequential data will be resampled to 16-day time series of spatial mean and standard deviation. Besides the remote sensing, weather, and soil features, two additional features are added to incorporate spatio-temporal information. These are year and geographic coordinates. A very important contributor to crop yield is cultivar, and the improvement in seed is the major cause of crop yield increase over the past twenty years. This technological factor can be approximately considered as a function of year.

Timeliness of Prediction

To evaluate prediction accuracy against timeliness, the following training and testing scenarios are adopted. Predictions are assumed to be made at four time points of a year: July, August, September, and October. At each time point of prediction, input observations should be taken before the specific time. The prediction in October is considered after-season prediction, which is also valuable since most official statistics are published at the beginning of the following year.

For data split and validation, training data is considered to be all the data from previous 10 years before the prediction time period. For example, to predict crop yield in August for 2012, training data is all the data from 2002 to 2011, and for each year take data before August. Such framework is also aimed to build a stable, robust, and operational approach for crop yield prediction in the future.

Reference


Irwin, S., D. Good, and D. Sanders. "Are USDA Corn Yield Forecasts Getting Better or Worse over Time?" farmdoc daily (4):166, Department of Agricultural and Consumer Economics, University of Illinois at Urbana-


