Computers and Electronics in Agriculture 153 (2018) 213–225

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield

Sami Khanal[a,*], John Fulton[b], Andrew Klopfenstein[b], Nathan Douridas[c], Scott Shearer[b]

[a] *Department of Food, Agricultural and Biological Engineering, Ohio State University, Wooster, OH 44691, USA*
[b] *Department of Food, Agricultural and Biological Engineering, Ohio State University, Columbus, OH 43210, USA*
[c] *Farm Science Review, Ohio State University, London, OH 43140, USA*

ARTICLE INFO

Keywords:
Remote sensing
Soil
DEM
Yield
Mapping

ABSTRACT

Widespread adoption of precision agriculture requires timely acquisition of low-cost, high quality soil and crop yield maps. Integration of remotely sensed data and machine learning algorithms offers cost-and time-effective approach for spatial prediction of soil properties and crop yield compared to conventional approaches. The objectives of this study were to: (i) evaluate the role of remotely sensed images; (ii) compare the performance of various machine learning algorithms; and (iii) identify the importance of remotely sensed image-derived variables, in spatial prediction of soil properties and corn yield. This study integrated field based data on five soil properties (i.e., soil organic matter (SOM), cation exchange capacity (CEC), magnesium (Mg), potassium (K), and pH) and yield monitor based corn yield data with multispectral aerial images and topographic data, both collected in 2013, from seven fields at the Molly Caren Farm near London, Ohio. Digital elevation model data, at a resolution of 1 m, was used to derive topographic properties of the fields. Multispectral images collected at bare-soil conditions, at a resolution 0.30 m, were used to derive soil and vegetation indices. Models developed for prediction of soil properties and corn yield using linear regression (LM) and five machine learning algorithms (i.e., Random Forest (RF); Neural Network (NN); Support Vector Machine (SVM) with radial and linear kernel functions; Gradient Boosting Model (GBM); and Cubist (CU)) were evaluated in terms of coefficient of determination (R^2) and root mean square error (RMSE). Machine learning algorithms were found to outperform LM algorithm for most of the times with a higher R^2 and lower RMSE. Based on models for seven fields, on average, NN provided the highest accuracy for SOM (R^2 = 0.64, RMSE = 0.44) and CEC (R^2 = 0.67, RMSE = 2.35); SVM for K (R^2 = 0.21, RMSE = 0.49) and Mg (R^2 = 0.22, RMSE = 4.57); and GBM for pH (R^2 = 0.15, RMSE = 0.62). For corn yield, RF consistently outperformed other models and provided higher accuracy (R^2 = 0.53, RMSE = 0.97). Soil and vegetation indices based on bare-soil imagery played a more significant role in demonstrating in-field variability of corn yield and soil properties than topographic variables. The accuracy of the models developed for prediction of soil properties and corn yield observed in this study suggested that the approach of integrating remotely sensed data and machine learning algorithms are promising for mapping soil properties and corn yield at a local scale, which can be useful in locating areas of potential concerns and implementing site-specific farming practices.

1. Introduction

Accurate and detailed information on soil properties and crop health is essential for optimization of farm management practices for sustainable production of agricultural goods and services (Souza et al., 2016; Yao et al., 2016), as well as for environmental modeling, and environmental risk assessment and management. High resolution maps of soil properties and crop yields enable producers and the agricultural community to identify in-field variability in soil and crop health and target areas within the field for soil fertility interventions, improved crop productivity, and better economic outcomes.

Traditional approaches for mapping soil properties and crop yield have mostly relied on field surveys and the use of costly equipment. Soil sampling and laboratory analyses are conducted for evaluating soil health, and harvester-mounted yield monitors are used for understanding the spatial variability in crop yield. These approaches however are time consuming and expensive, especially when mapping needs to be done at regional, national, and global scales (Mulder et al., 2011;

* Corresponding author.
E-mail address: Khanal.3@osu.edu (S. Khanal).

https://doi.org/10.1016/j.compag.2018.07.016
Received 9 January 2018; Received in revised form 17 April 2018; Accepted 8 July 2018

Table 1
Basic characteristics of the fields studied, including field size, slope, dominant soil map unit, dominant soil order, number of soil samples, and field management practices.

Field	Size (ha)	Slope (%)	Soil map unit	Dominant soil order	Sample number	Tillage	Crop rotation
1B	11	4.37	Ochraqualfs (40.7%), Argiaquolls (31%), Epiaqualfs (18%), Argiudolls (10.3%)	Alfisols	27	NT	C-C-S
1C	5.3	5.86	Ochraqualfs (74%), Argiaquolls (26%)	Alfisols	17	CT	C-S-C
1D	6.5	4.35	Ochraqualfs (94.8%), Argiaquolls (5.2%)	Alfisols	20	NT	C-S-C
9A	13.3	5.7	Ochraqualfs (58%), Argiaquolls (42%)	Alfisols	39	CT	C-S-C
12D	17.5	4.98	Argiaquolls (46%), Hapludalfs (27.9%); Ochraqualfs (23.8%)	Mollisols	49	CT	W-S-C
MISD	12	9.26	Ochraqualfs (82.5%), Argiaquolls (18.5%)	Alfisols	36	CT	S-W-C
PENIN	3.8	9.6	Ochraqualfs (98%)	Alfisols	12	NT	C-C-S

Tillage: NT – No Till; CT – conventional tillage (i.e., field cultivator was used prior to planting the crop). ***Crop Rotation***: C- Corn; S- Soybean; W-Wheat.

Yang et al., 2014). Furthermore, these approaches have several limitations. For example, yield monitor based data can only be collected at harvest and, thus, cannot be used for in-season crop management. Also, these data are spatially coarse and fail to capture in-field variability in soil and crop health (Souza et al., 2016).

Remotely sensed images have the potential to overcome the limitations of traditional approaches and improve the spatial coverage of soil and crop yield data (Peng et al., 2015; Stevens et al., 2013; Yao et al., 2016). Studies have demonstrated that many soil properties can be estimated by integrating georeferenced field collected soil and crop data with spectral properties of soil acquired by sensors onboard satellite and aircrafts. Dobos et al. (2001) found the Advanced Very High Resolution Radiometer (AVHRR) satellite data and DEM derived terrain variables to be powerful in characterizing soil-forming environments and delineation of soil patterns on a regional scale. Scudiero et al. (2014) found multi-year spectral reflectance data from the Landsat to be a reliable indicator of soil salinity in the western San Joaquin Valley in California, USA. Several studies have also been conducted focusing on crop yield mapping by integrating remotely sensed images acquired from satellite (Lobell et al., 2015), aircraft (Yang et al., 2014), and unmanned aerial vehicles (Geipel et al., 2014; Shi et al., 2016).

Despite prior efforts, further exploration on the application of remotely sensed data for mapping of soil properties and crop yield is needed. The success in prediction and mapping of soil properties, and crop health and yield using remotely sensed data to a large extent depends on the availability, quality, and timing of remotely sensed data collection (Blasch et al., 2015), as well as the approaches used for model development (Forkuor et al., 2017; Morellos et al., 2016). Prior studies have mostly focused on estimating crop yield and soil properties at regional scales rather than for individual fields (Lobell et al., 2015). These studies used satellite acquired remotely sensed images with coarse spatial resolution. Mapping of soil properties and crop yield at coarse resolution is of limited use for resource assessment and management at a field scale; whereas, maps at high resolution can help the agricultural and environmental community to cost-effectively detect and characterize the extent of soil and crop health issues. This information can be used for prescription-based farming that help improve economic outcome and environmental footprints associated with agricultural practices.

A linear regression algorithm is the most commonly used approach to estimate crop yield and soil properties (Geipel et al., 2014; Lobell et al., 2015). However, it has limitations in handling non-linear relationships between response and predictor variables that usually exist in heterogeneous agricultural landscapes. There are several machine learning algorithms that can overcome this limitation, and provide better prediction of soil variables and crop yield. However, comparisons of the traditional linear regression algorithm to machine learning algorithms for prediction of soil properties and crop yield are limited. In addition to understanding the performance of various models in mapping soil properties and crop yield, there is a need to identify the relative importance of variables for enhancing the predictive ability of the models.

The objectives of this study were to: (i) examine the role of remotely sensed images; (ii) evaluate the performance of linear regression and machine learning algorithms; and (iii) identify the importance of remotely sensed image-derived variables, for prediction and mapping of soil properties and corn yield. Seven statistical models were developed for predicting corn yield and soil properties. Soil properties examined in this study included soil organic matter (SOM), cation exchange capacity (CEC), potassium (K), magnesium (K), and pH. Prior studies (Forkuor et al., 2017; Morellos et al., 2016) have used remotely sensed data for mapping of soil properties; however, this is to our knowledge the first evaluation of remotely sensed images of bare soil surface at a spatial resolution < 1 m from multiple fields for prediction and mapping of both soil properties and corn yield.

2. Materials and methods

2.1. Study area

Fields examined in this study are located in the northwest part (83°26′14.3″–83°26′49.24″W, 39°56′37.82″–39°57′28.7″N) of Madison County, Ohio, USA. The dominant soil types in these fields are Ochraqualfs (Crosby-Lewisburg Complex), Argiaquolls (Kokomo Silty Clay Loam, Westland silty clay loam), and Hapludalfs (Miamian Silt Loam, Eldean silt loam, Thackery variant silt loam) (Table 1). These fields are gently rolling, with the mean slope ranging from 4.35 to 9.26%. The average elevation of the fields is 311 m. The mean annual rainfall (1981–2016) is 998 mm with approximately 58% of annual rainfall occurring between April and September. The mean annual temperature is 10.9 °C, with daily temperatures ranging from −6.7 (minimum) to 29.2 °C (maximum).

A strong spatial variability in soil properties was observed in the study area. Soil properties were characterized by large range and high standard deviation, with SOM in the range of 1.2–4.9 (%), CEC of 6–27.3 (meq/100 g), K of 1.2–5.9 (%), Mg of 10.2–36.7 (%), and pH of 5–78 (Table 2).

2.2. Data

2.2.1. Soil and crop data

A total of 200 soil samples were collected from seven bare fields (Table 1) in October 1, 2013. In each field, samples were taken at a depth of 18 cm on 1-acre intervals. The samples were air-dried at 49 °C (120 °F) for 24 h, sieved, and sent to the Spectrum Analytic lab (Spectrum Analytic, 2017) for soil analyses. As field 12D has very different soil map units compared to six other fields (Table 1), soil samples were classified into two dominant soil orders (Alfisols and Mollisols), and a "group" was introduced as an independent variable for model development.

Corn yield data were available for only one field (i.e.,12D), and thus, the models for corn yield prediction were focused on this field only. Corn yield data were recorded by a John Deere yield monitoring system during harvest. The yield monitor was calibrated before and