Characterizing the spatial variability of soil properties and crop yield using high-resolution remote sensing image and ground-based data

Sami Khanal[1], John Fulton[1*], Andrew Klopfenstein[1], Nathan Douridas[2], and Scott Shearer[1]

1 Department of Food, Agricultural and Biological Engineering, Ohio State University

2 Farm Science Review, Ohio State University

Abstract

Site-specific management practices offer a strategy to optimize agronomic inputs if the causes for spatial and temporal variability in crop yield are understood, and can be related to one or more field properties. The objectives of this study are to: (i) determine the suitable methodology for predicting soil properties and crop yield at high-spatial-resolution by evaluating various machine-learning algorithms, and (ii) identify the relationships between crop yield, soil and topographic properties. The analyses will be based on four soil properties (soil organic carbon, potassium, soil cation exchange capacity and magnesium), and maize (*Zea mays* L.) yield data from a maize-soybean (*Glycine max*) rotation field in central Ohio, USA. Topographic properties including slope, aspect, roughness, terrain ruggedness index, topographic position index, and flow direction were derived from a high-spatial-resolution (1 m) digital elevation model data. Soil and vegetation indices were derived from a high-spatial-resolution (0.30 m) multispectral bare-soil aerial image. Soil properties were estimated at high-spatial-resolution by integrating field-collected soil samples, topographic properties, and spectral information (individual spectral bands and indices) derived from multispectral aerial image. Continuous maize yield estimates were then predicted by integrating aforementioned data. Multiple linear regression (LM) and two machine-learning algorithms, including random forest (RF) and neural network (NN), were investigated to evaluate and identify the method that best characterized the relationships between bare-soil multispectral image, soil properties and crop yield. Comparison of models for estimating soil properties and crop yield demonstrated that soil organic carbon and crop yield can be predicted using bare-soil multispectral image with higher accuracy using the RF model. The LM model predicted soil cation exchange capacity and magnesium with higher accuracy, and NN performed better in estimating potassium. Based on the analyses of importance of variables across the three models for maize yield prediction, variables including red and green spectral bands, and soil indices derived from bare-soil multispectral image, and cation exchange capacity were found to be the five most significant predictors influencing maize yield estimates.

Keywords: Remote sensing, Yield, Variability, Soil, Topography

Background

Site-specific farming, also known as precision agriculture (PA), promises to improve farm productivity and resource efficiency, and reduce costs and negative environmental footprints. Although PA has been an agricultural goal for a very long time, it gained new emphasis due to growing concerns about environmental degradations that have resulted from intensive agricultural practices. PA aims to optimize soil and crop management practices by applying inputs in accordance with site-specific requirements of a specific soil and crop, which vary in time and space (Moshia et al. 2014).

Accurate and detailed information on soil properties and crop yield is essential for PA (Al-Gaadi et al. 2016; Gelder et al. 2011; Yao et al. 2016) as well as for environmental modelling (Miller and White 1998) and risk assessment. High-spatial-resolution information on soil and crop can help producers and agricultural community to better characterize soil and crop health and targets areas within the field for soil fertility interventions and ultimately, improved crop productivity and a better economic outcome. However, traditional soil and crop mapping approaches mostly rely on field surveys or equipment that are time consuming and expensive if mapping needs to be done at regional, national and global scales (Mulder et al. 2011; Yang et al. 2013). Although yield monitors are commercially available, many crop harvesters are not equipped with them. Also, yield monitor data can only be collected at harvest and used for after-season management. Moreover, the data based on soil sampling and yield monitor are spatially coarse, and thus have limitations.

Techniques that combine georeferenced soil data and remotely sensed imageries have potential to overcome the limitations of traditional approaches and improving the detail and spatial coverage of soil properties and crop yield patterns (Gelder et al. 2011; Yang et al. 2013; Yao et al. 2016). Studies (Barnes et al. 2000; Chen et al. 2005, 2008; Gomez et al. 2008; Hahn and Gloaguen 2008; Minasny et al. 2008) have demonstrated the correlation between soil properties and spectral reflectance, and the importance of remote sensing data in mapping soil properties such as sand, silt, clay, and soil organic carbon (SOC). Similarly, studies (Dobermann and Ping 2004; Yang et al. 2007) have also used satellite and airborne multispectral images for mapping in-field crop variation for preseason, within-season and after season management.

Access, quality and feasibility of imagery for agriculture applications has greatly advanced in recent years in the US. Despite prior efforts, further exploration of the applications of remote sensing data for soil and crop yield mapping is needed. Few studies exist, if any, that have focused on the use of remote sensing data at a spatial resolution lower than 1 m for soil mapping, and have integrated high-spatial-resolution soil information for crop yield mapping. Traditionally, multiple linear regression models have been used in mapping soil properties and crop yield. Only few studies have explored the use of machine-learning algorithms in estimation of soil properties and crop yield. Thus, the objectives of this study are to (i) determine the suitable statistical methods for predicting soil properties and crop yield, and (ii) identify the relationships between crop yield, soil and topographic properties.

Methods

Study area

Field examined in this study is located in the northwest part (-83° 25' 17.57", 39° 57' 11.12’’N) of Madison County, Ohio, USA. The dominant soil types in this fields are Argiaquolls (Kokomo Silty Clay Loam, Westland silty clay loam), Hapludalfs (Miamian Silt Loam, Eldean silt loam, Thackery variant silt loam), and Ochraqualfs (Crosby-Lewisburg Complex). This field is gently rolling, with an average slope of 4.96%. The average elevation of the fields is 307 meter. The mean annual rainfall (1981-2016) is 998 mm with approximately 58% of annual rainfall occurring between April and September. Mean annual temperature is 10.9 °C with daily temperature ranges between -6.7 (minimum) to 29.2° C (maximum).

Soil and crop data

A total of 49 soil samples were taken from the study region on 1 October 2013. Samples were taken at a depth of 18 cm on one-acre interval and dried in a forced air-drying room at 49°C (120 F) for 24 hours and ground. Soil properties including cation exchange capacity (CEC), potassium (K) and magnesium (Mg) were estimated following the approach described in Dellavalle (1992) with the instrument Thermo Scientific iCAP 6200 ICP-OES. Soil organic carbon was estimated with the Costech elemental combustion analyzer. Maize (Zea mays L.) yield estimates were based on yield monitor of a John Deere harvester. The yield monitor was calibrated before and during the harvest to minimize the potential error in yield estimates. As yield data near the corner of the field was found to have some errors due to change in machine orientation (Lyle et al. 2014)**,** an inner buffer with a distance of 20 m from the field boundary was created to exclude yield data near the corner of the field from the analyses.

Remote sensing data

A multispectral image of the study region was obtained under the Ohio Statewide Imagery Program in May of 2013. The ground was bare during the time of image acquisition. This image has visible (red, green, blue) and near-infrared bands at 0.30 m spatial resolution, and it was collected with the Leica ADS80 digital camera, and rectified using LiDAR data. In addition to the spectral values, recorded as digital numbers in each of the visual and NIR wavebands of the multispectral image, six soil and vegetation indices (Table 1) that were found useful in digital soil mapping (Ray et al. 2004) were calculated from the combination of wavebands in the multispectral image. A 0.76 m resolution DEM data available from the Ohio Geographically Referenced Information Program was used to extract terrain variables. Prior to the calculation of terrain variables for analyses, DEM was pre-processed to generate a depression free DEM.

Table 1. Soil and vegetation indices considered for the analyses

Indices	Formula	Index Property	Reference
Brightness Index (BI)		Average reflectance magnitude	Ray et al., 2004
Saturation Index (SI)		Spectral Slope	Ray et al., 2004
Hue Index (HI)		Primary Color	Ray et al., 2004
Coloration Index (CI)		Soil Color	Ray et al., 2004
Redness Index (RI)		Hematite Content	Ray et al., 2004
Normalized Difference Vegetation Index (NDVI)		Health and amount of vegetation	Ray et al., 2004

To ensure the proper integration with multispectral image, DEM was resampled to 0.30 m resolution (the resolution of the multispectral images) using the bilinear interpolation method. A low-pass filter with a 5 by 5 cell mask was applied to each band of the multispectral image, and DEM to minimize the potential variance among the image pixels that might have been introduced by microtopography, image processing and scanning (Chen et al., 2005; Hively et al., 2011). Spectral values in individual wavebands, spectral indices, and terrain properties were extracted to soil sampling locations to develop the statistical models for estimating soil properties. Table 2 lists the terrain variables used in the analyses. To our understanding, DEM and the multispectral image used in this study are the highest resolution datasets ever used in mapping of soil property. A rectangle with the length equal to harvester swath width (6 m) and width equal to the distance between two reported maize yield locations (2 m) was drawn around each reported maize yield locations to extract information from images and relate with maize yield. This is done because the size of the area represented by a yield pixel/location in a yield monitor is proportional to speed and header width of the harvester. Image derived information were extracted at these polygons using the zonal statistics function of the Spatial Analyst tool in the ArcGIS.

Table 2. Terrain variables considered in the study

Parameters	Definition	Units	References
Slope	Inclination of the land surface from the horizontal	Degree	(Allen et al. 2014)
Aspect	Direction the slope faces	Degree	(Davy and Koen 2014)
Roughness	Difference between maximum and minimum elevation	-	
Terrain Ruggedness Index (TRI)	Amount of elevation difference between neighbouring areas	-	
Topographic Position Index (TPI)	Measure of where a location is in the overall landscape	-	(Wilson et al. 2007)
Flow Direction	Path of water flow	-	(Kitchingman and Lai 2004)

Statistical analyses

Three statistical models, including multiple linear regression (LM), random forest (RF) and neural network (NN), were considered for predicting soil properties and maize yield. The performance of these three models was then compared to determine the best model. All the statistical analyses were performed in the R software. The statistical package "caret" was used for the model development. To provide an unbiased sense of model effectiveness, the total data was split into training and test datasets at 4:1 ratio, where the training set was used to create the model and the test dataset was used to evaluate the model performance. Mean SOC, CEC, K and Mg values between the training and test datasets are ensured to be similar so that calibration models are well trained to predict the range of soil properties of the test dataset.

For predicting soil properties, soil variables including SOC, CEC, K and Mg were the dependent variables, and spectral and terrain variables were the independent variables. For predicting maize yield, observed maize yield was the dependent variable, and all other variables were the independent variables. The functions “lm”, “rf” and “nn” in the R software were used for developing LM, RF, and NN models, respectively. One soil property was modeled at a time as the dependent variable against all predictor variables. For each model, the adjusted R^2 and residual standard error were considered. In addition, the predictors with significance ($p<0.01$) were noted. A stepwise regression analyses was conducted to address a problem of multicollinearity (i.e., significant correlation between the predictors) in the LM model. Stepwise regression identifies a subset of predictors based on their statistical significance using stepwise, forward selection, or backward elimination. For stepwise regression, “stepAIC” function in the “MASS” package of the R software (Kuhn and Johnson 2013) was used. Table 3 provides the number of predictors that were eventually used in the LM model for each soil property. The same set of predictors that were determined after stepwise regression were maintained across all models to provide comparison with other models. Prior to running the models, all the predictor variables were centered and scaled to ensure all were on the same scale.

The parameters required for tuning the RF model such as the number of predictors that are randomly sampled as candidates for each split (mtry) and the number of trees to grow in the forest (ntree) were set by using the grid search method in the “caret” package in R using tenfold cross-validation with five repetitions. Similarly, the required parameters for NN model fitting (i.e., hidden layer, decay rate) were set using the tenfold cross-validation with five repetitions also with the “caret” package (Kuhn and Johnson 2013). The performance of the three models investigated was assessed based on (1) model generated accuracy statistics based on training dataset used for model development, and (2) test dataset for model validation. The R^2 and RMSE derived from the three models for the respective soil parameters and maize yield were compared. A variable importance measure was estimated to understand the relative importance of predictors to the outcome of various models. A varImp function in package “caret” in the R software (Kuhn 2017) was used for this purpose.

Table 3. Number of explanatory variables used in the modeling of each soil parameter

Parameter	Spectral Variables	Terrain Variables
Soil Organic Carbon (SOC)	Red, green, blue, BI, SI, HI, CI	Rough
Cation Exchange Capacity (CEC)	Green, SI, HI, CI,BI	
Potassium (K)	Green, blue, SI, CI, BI, RI,	TRI, Slope
Magnesium (Mg)	SI, HI, CI	Aspect

Results

Model performance in estimating soil properties

The performance of the models for estimating soil properties including SOC, CEC, K and Mg suggested that soil properties can be estimated using high-resolution bare-soil multispectral image with relatively higher accuracy. In this study, CEC was estimated with higher accuracy followed by K, Mg and SOC. Assessment of the model performance based on cross-validation of the training set resulted in R^2 that ranged between 0.66 and 0.78 for CEC, 0.47 and 0.53 for K, 0.38 and 0.51 for Mg, and 0.36 and 0.49 for SOC (Table 4). Except for few models for K and SOC estimations, the performance of the majority of models were found poor (i.e., lower R^2 and higher RMSE) with the test dataset. This is probably due to the larger number of training dataset compared to test dataset, which confirms the importance of obtaining representative training set to produce reliable predictions. While comparing the performance of three models for both training and test datasets, none of the models performed consistently superior over the other for the prediction of four soil properties. For instance, with the test dataset, RF model for SOC performed better resulting in higher R^2 and lower RMSE than LM and NN models. LM performed marginally better in estimating CEC and Mg, and NN performed better in estimating K. However, during the model validation with the test dataset, performance of NN and RF models were superior for SOC and CEC, respectively. With the test dataset of K and Mg, LM performed better than the other models.

Table 4. Model performance for estimating soil properties

	SOC		CEC		K		Mg	
Models	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
LM	0.38(0.0)	0.41(0.61)	0.78(0.59)	1.70(3.15)	0.47(0.56)	0.20(1.63)	0.51(0.35)	3.58(3.92)
RF	0.49(0.21)	0.39(0.47)	0.72(0.66)	1.89(3.07)	0.50(0.13)	0.58(0.87)	0.48(0.09)	3.79(4.5)
NN	0.36(0.52)	0.43(0.52)	0.66(0.59)	2.14(3.20)	0.53(0.34)	0.55(0.76)	0.38(0.32)	4.34(3.72)

*model performance based on test datasets are provided within the parenthesis

Model performance in estimating maize yield

Table 5 shows the average R^2 and RMSE of three models investigated for maize yield prediction based on cross-validation with training dataset and validation with test dataset. R^2 ranged between 0.29 and 0.46 during the cross-validation, and 0.30 and 0.46 during the validation phases. RF model performed better (i.e., higher R^2 and lower RMSE) than the other models for both the cross-validation and validation phases.

Table 5. Model performance for estimating maize yield

Models	Cross-Validation with Training Dataset		Validation with Test Dataset	
	R^2	RMSE	R^2	RMSE
LM	0.29	1.30	0.30	1.29
RF	0.46	1.13	0.46	1.13
NN	0.27	1.31	0.30	1.28

* LM – Linear Model; RF – Random Forest, NN – Neural Network

Relationships between maize yield, soil and topographic properties

Of the twenty variables considered for model development for maize yield prediction, all variables except BI, NDVI, TPI and FlowDir were found to have significant influence on maize yield. Based on the analyses of importance of selected variables for maize yield prediction across three models, variables including red, green, SI, CI, RI and CEC were found to be the five most significant predictors for three models (Table 6). Based on RF model, relative position of a surface with respect to its local neighbourhood indicated by parameter “Rough” was found to be the most important predictor contributing to the accuracy of the maize prediction. Soil spectral characteristics indicated by indices such as SI, RI and CI were also found to influence maize yield estimates. Of the soil properties, CEC was found to influence yield the most.

Table 6. First five predictors that were highly significant across models

Models	Predictors*
LM	RI, Mg, red, green, CI
RF	Rough, SI, RI, CEC, CI
NN	red, green, blue, SI, CEC

*predictors in the left are as ranked as first and the right are ranked as the fifth most important predictors

Mapping the spatial distribution of maize yield

Except at the corners of the field, the spatial distribution of predicted maize yield was found to be similar to the pattern of observed maize yield estimates (Figure 1). Except for the 2% data in the corner of the fields, the RF model predicted maize yield reasonably well with average difference of -0.48% (±6.9 % standard deviation) between predicted and observed yield. While comparing the predicted maize yield with observed yield estimates, similar spatial distributions of hot and cold spots of maize yield were observed.

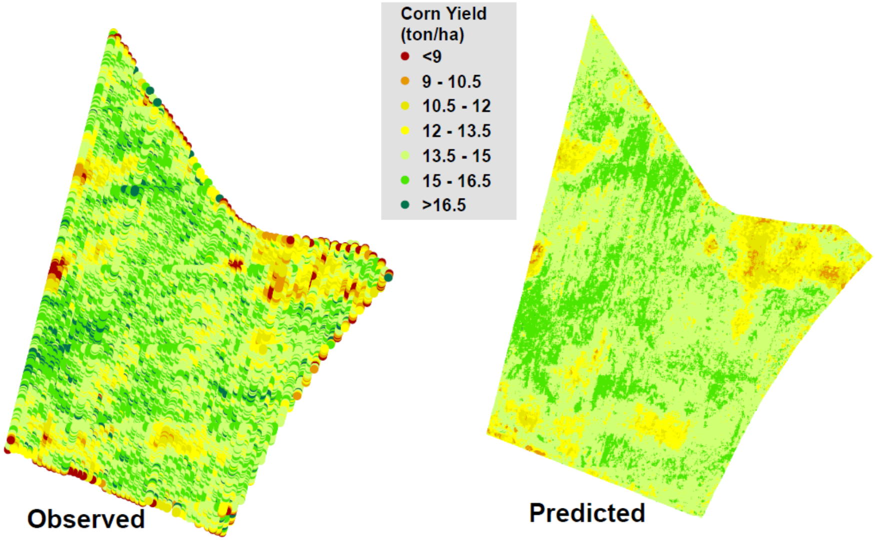

Figure 1. Maps of observed and predicted maize yield (ton/hectare)

Discussion

The models were calibrated and tested with data collected from one field for one year. This means that the models developed in this study cannot be generalized for the prediction of same soil parameters, and maize yield in other soil types and geographic regions. The generalization of the models is strongly dependent upon the variability of the training and test datasets. Thus, there is a need to use data from other fields with varying management practices and soil types, and from multiple years. Nevertheless, the analyses presented in this study demonstrate that remote sensing data and machine-learning approaches could be adopted for cost-effective prediction of high-spatial-resolution soil and crop yield mapping.

Conclusion

Results indicated that readily available remote sensing images can be used to generate continuous high-spatial-resolution and accurate estimates of soil properties and crop yield. For the study site, the RF performed better in estimating SOC and maize yield. While the LM model performed better in estimating CEC and Mg, the NN performed better in K estimation. Variables such as surface roughness followed by soil indices (SI, RI) and CEC were found to contribute the most to prediction accuracy of maize yield. The use of remote sensing data for assessing soil and crop health help reduce soil sampling efforts and the associated costs. Further, these information help assess fertility requirements, crop yield trends, and locate poor and high productivity areas within the field to further advance site-specific farming in this region.

Acknowledgments

Partial funding for this project was provided by the Field to Faucet program under the College of Food, Agricultural, and Environmental Science at the Ohio State University under Grants No. F2F-000004. We also appreciate the support of the Farm Science Review and Molly Caren Ag Center where this study was conducted.

References

Al-Gaadi KA, Hassaballa AA, Tola E, Kayad AG, Madugundu R, Alblewi B, Assiri F 2016. Prediction of potato crop yield using precision agriculture techniques. PloS one 11(9): e0162219.

Allen DE, Pringle MJ, Bray S, Hall TJ, O'Reagain PO, Phelps D et al. 2014. What determines soil organic carbon stocks in the grazing lands of north-eastern Australia? Soil Research 51(8): 695–706.

Barnes EM, Baker MG, others 2000. Multispectral data for mapping soil texture: possibilities and limitations. Applied Engineering in Agriculture, 16(6): 731–746.

Chen F, Kissel DE, West LT, Adkins W, Rickman D, Luvall JC 2008. Mapping Soil Organic Carbon Concentration for Multiple Fields with Image Similarity Analysis. Soil Science Society of America Journal 72(1): 186. DOI: 10.2136/sssaj2007.0028.

Chen F, Kissel DE, West LT, Rickman D, Luvall JC, Adkins W 2005. Mapping Surface Soil Organic Carbon for Crop Fields with Remote Sensing. Journal of Soil and Water Conservation 60(1): 51–57.

Davy MC, Koen TB 2014. Variations in soil organic carbon for two soil types and six land uses in the Murray Catchment, New South Wales, Australia. Soil research 51(8): 631–644.

Dellavalle NB 1992. Handbook on reference methods for soil analysis. Soil and Plant Analysis Council. Inc., Lincoln, Nebraska.

Dobermann A, Ping JL 2004. Geostatistical integration of yield monitor data and remote sensing improves yield maps. Agronomy journal 96(1): 285–297.

Gelder BK, Anex RP, Kaspar TC, Sauer TJ, Karlen DL 2011. Estimating Soil Organic Carbon in Central Iowa Using Aerial Imagery and Soil Surveys. Soil Science Society of America Journal 75(5): 1821. DOI:10.2136/sssaj2010.0260.

Gomez C, Rossel RAV, McBratney AB 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma 146(3): 403–411.

Hahn C, Gloaguen R 2008. Estimation of soil types by non linear analysis of remote sensing data. Nonlinear Processes in Geophysics 15(1): 115–126.

Kitchingman A, Lai S 2004. Inferences on potential seamount locations from mid-resolution bathymetric data. Focus 32(64): 128.

Kuhn M 2017. CARET: Classification and Regression Training.

Kuhn M, Johnson K 2013. Applied Predictive Modeling. DOI: 10.1007/978-1-4614-6849-3.

Lyle G, Bryan BA, Ostendorf B 2014. Post-processing methods to eliminate erroneous grain yield measurements: review and directions for future development. Precision agriculture 15(4): 377–402.

Miller DA, White RA 1998. A Conterminous United States Multilayer Soil Characteristics Dataset for Regional Climate and Hydrology Modeling. Earth Interactions 2(1): 2–2. DOI: 10.1175/1087-3562(1998)002<0002:CUSMS>2.0.CO;2

Minasny B, McBratney AB, Tranter G, Murphy BW 2008. Using soil knowledge for the evaluation of mid-infrared diffuse reflectance spectroscopy for predicting soil physical and mechanical properties. European Journal of Soil Science 59(5): 960–971.

Moshia ME, Khosla R, Longchamps L, Reich R, Davis JG, Westfall DG 2014. Precision manure management across site-specific management zones: Grain yield and economic analysis. Agronomy Journal 106(6): 2146–2156.

Mulder VL, De Bruin S, Schaepman ME, Mayr TR 2011. The use of remote sensing in soil and terrain mapping-A review. Geoderma 162(1): 1–19.

Ray SS, Singh JP, Das G, Panigrahy S 2004. Use of high resolution remote sensing data for generating site-specific soil management plan. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci 35(B7): 127–132.

Wilson MFJ, O’Connell B, Brown C, Guinan JC, Grehan AJ 2007. Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope. Marine Geodesy 30(1–2): 3–35.

Yang C, Everitt JH, Bradford JM 2007. Airborne hyperspectral imagery and linear spectral unmixing for mapping variation in crop yield. Precision Agriculture 8(6): 279–296.

Yang C, Everitt JH, Du Q, Luo B, Chanussot J 2013. Using high-resolution airborne and satellite imagery to assess crop growth and yield variability for precision agriculture. Proceedings of the IEEE 101(3): 582–592.

Yao RJ, Yang JS, Wu DH, Xie WP, Gao P, Wang XP 2016. Characterizing Spatial--Temporal Changes of Soil and Crop Parameters for Precision Management in a Coastal Rainfed Agroecosystem. Agronomy Journal 108(6): 2462–2477.