Statistical and machine learning methods for crop yield prediction in the context of precision agriculture

Hannah Burdett1 · Christopher Wellen1

Accepted: 3 March 2022 / Published online: 30 March 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

It is of critical importance to understand the relationships between crop yield, soil properties and topographic characteristics for agricultural management. This study’s objective was to compare techniques to quantify the relationship between soil and topographic characteristics for predicting crop yield using high-resolution data and analytical techniques. The study was conducted on a multiple field dataset located in Southwestern Ontario, Canada, where few studies have assessed the impact of applications for precision agriculture and machine learning (ML) to the soil property-yield relationship in this region. The dataset included 145,500 observations of corn and soybean yield, topographic and soil nutrient characteristics. The attributes considered for this study included pH, soil organic matter (OM) content, cation exchange capacity (CEC), soil test phosphorus, zinc (Zn), potassium (K), elevation and topographic wetness index. Multiple linear regression (MLR), artificial neural networks, decision trees and random forests were compared to identify methods able to relate soil properties and crop yields on a subfield scale (2 m). Random forests were the most successful at predicting yield with an $R^2$ value of 0.85 for corn and 0.94 for soybeans. MLR was the least successful with an $R^2$ of 0.40 for corn and 0.45 for soybeans. Cross-validation experiments showed that random forest models in most cases could predict low- and high-yield areas from fields excluded from training datasets, but this was not possible in all cases. Techniques tested the models and identified significant soil and topographic attributes when predicting yield, though the identification was subject to some uncertainty. These results suggest that ML techniques might be used to predict high yield areas of fields without existing yield maps, if those fields have similar relationships of soil properties to yield.

Keywords Spatial cross-validation · Variable importance · Spatial data mining · Regression · Yield · Phosphorus

Hannah Burdett
hburdett@uwaterloo.ca

1 Department of Geography, Faculty of Arts, Ryerson University, Toronto, ON, Canada
Introduction

Agronomic scientists have conducted extensive research to map, monitor, analyze and manage yield variability to optimize crop yield (Miao et al., 2006). One technique that assists with crop management is the use of crop yield predictions. Crop yield predictions are used by crop managers to reduce losses by recognizing areas or factors that may cause adverse growing conditions, such as crop responses under climatic stress or deficiencies in nutrients. Additionally, crop yield predictions could be used to evaluate the optimum growing conditions so that fields may achieve their full growth potential (Dahikar & Rode, 2014). Crop yield depends on many inter-related factors such as elevation, cation exchange capacity (CEC) and soil nutrients (Miao et al., 2006). Predicting crop yield can be difficult as there is an increasing recognition that relationships between ecological drivers and their responses are commonly complex and non-linear (D’Amario et al., 2019; Gonzalez-Sanchez et al., 2014).

Several techniques have been used to understand the relationships between crop yield and soil or landscape properties (Miao et al., 2006). Statistical models such as multiple linear regression (MLR) have been widely applied in the context of precision agriculture (Drummond et al., 2003; Kravchenko & Bullock, 2000; Sudduth et al., 1996). However, results from MLR are often not satisfactory, as MLR is limited to describing linear relationships between crop parameters and site variables and the results are potentially misleading when these relationships are not linear (Drummond et al., 2003; Liu et al., 2001). For instance, Sudduth et al. (1996) used linear techniques on a dataset consisting of several site-years of topographic, soil and yield data. They found that linear methods generally failed to produce reasonable approximations of spatial yield variability, even with sub-field regions thought to be reasonably homogenous.

Machine learning (ML) techniques have been applied through various agricultural systems over the past decade to provide more accurate solutions, primarily because of their ability to handle highly complex non-linear agricultural problems (Pantazi et al., 2016; Tantalaki et al., 2019). Unlike traditional statistical methods, ML does not make assumptions about the correct structure of the data model, such as the functional form, probability distribution or smoothness (Khazaei et al., 2008; Mittal & Zhang, 2000; Seyhan et al., 2005). Instead, ML techniques have the ability to learn the relationship between dependent and independent variables through the data (Mittal & Zhang, 2000). ML techniques are based on semi-parametric and non-parametric structures, with validation relying on prediction accuracy (Gonzalez-Sanchez et al., 2014). A non-parametric approach does not require any prior assumptions about the form of probability distributions that the class and other attributes satisfy. In the context of crop yield prediction with multiple inputs, applications of such techniques have produced higher accuracy models (Dahikar & Rode, 2014).

Some comparisons among linear and ML techniques for crop yield prediction have been made. Gonzalez-Sanchez et al. (2014) compared MLR, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression and k-nearest neighbor for fields in Mexico. The dataset included records from the fall-winter season in the years 1998–2006 and included a total of 6217 observations. However, this study focused primarily on the effects of temperature, solar radiation, rainfall, irrigation water depth, season duration and planting area to predict crop yield and excluded soil nutrients and topographical characteristics. It has been established that soil properties, topographic characteristics and crop yield exhibit spatial variability within agricultural fields (Corwin & Lesch, 2003; Jung et al., 2006; Metwally et al., 2019). For instance, Kravchenko and Bullock (2000)
conducted a study on eight fields with soil data from 1994 to 1997 in central Illinois and eastern Indiana, USA to evaluate how useful topographical information and soil properties can explain yield variability. The cumulative effect of the topographical features explained about 20% of the yield variability, while soil properties explained approximately 30% of the yield variability. Thus, spatial variability of soil and topographic properties can account for approximately 50% of agricultural production (Dahikar & Rode, 2014).

Previous research comparing linear and ML techniques often only provide insight into the efficacy of the model’s ability to predict yield (Drummond et al., 2003). However, cross-validation techniques in ML studies are often not utilized or are limited to resampling procedures such as split-sample procedures. For instance, Drummond et al. (2003) compared artificial neural networks (ANNs), stepwise MLR and projection pursuit regression. Through this comparison, Drummond et al. (2003) identified ANNs as the most effective method for crop yield prediction and applied a fivefold cross-validation technique to avoid overfitting. In the study presented, cross-validation techniques were applied to test the performance of the model, gain insight into the effect of soil and topographic attributes on yield and to test the relationships between fields.

Since the early 2000s, the runoff from agricultural nutrients and sediments has led to the re-eutrophication of lakes and impaired stream health in the Great Lakes Basin, USA, Canada (Kerr et al., 2016; Mohamed et al., 2019). The 4R nutrient stewardship program is promoted through a partnership with the private sector of agriculture nutrient service providers. It is focused on helping growers in the Western Lake Erie Basin use the right fertilizer source, “at the Right rate, at the Right time and with the Right placement” (Vollmer-Sanders et al., 2016), with the intention that this approach will optimize nutrient use while reducing nutrient losses to aquatic systems. With this concept in mind and some cases for efficiency’s sake, farmers in Southern Ontario, Ontario have been practicing precision agriculture management strategies. Precision agriculture (PA) utilizes variability in soil and topographic properties to manage fields through site-specific management strategies (International Society of Precision Agriculture, 2019). Fields with a higher degree of spatial variability are likely to benefit from such crop management strategies (Mzuku et al., 2005). PA uses a wide range of technologies to gather, process and examine data to guide targeted actions that advance the efficiency, sustainability and productivity of agricultural practices (Tey & Brindal, 2012). This management strategy has the potential to drive a new wave of increased agricultural productivity as well as contribute to the environmental sustainability of farming practices (Robertson et al., 2007). However, few studies have looked into the effect of such practices in Southern Ontario.

The objectives of this study were to (1) evaluate the predictive ability of MLR, ANNs, random forests (RF) and decision trees (DT) to identify which techniques provide the most accurate predictions of point-scale yield, (2) identify which fields in operation have similar relationships of soil properties to yield and which fields differ, and (3) identify which variables may be the most important for yield.

Materials and methods

Study area

Data was collected from 17 fields located in Southwestern Ontario, Canada. All the fields are owned by the same farmer and undergo similar farm management practices.
The fields range from 2 to 32 ha in size. Cash crops were grown in a rotation of corn and soybeans, with no animals present and no manure used. The soil in this area is comprised of Brookston clay soil with subsoil claypan horizon(s) varying between silty clay loams, silty clay, clay loam or clay (Richards et al., 1949). The topography of Southwestern Ontario is primarily flat, with slopes less than 1% (Frank & Ripley, 1977). Southwestern Ontario generally has a drier, warmer climate with sufficient soil fertility; thus, it is considered ideal for agricultural practices (Tan & Reynolds, 2003). Soils throughout southwestern Ontario generally in the study site specifically are tile drained (Nelligan et al., 2021).

Data

Crop yield and soil nutrients data

The yield and crop data were collected by professional agronomists. Crop yield measurements for this study were obtained using a full-size combine equipped with a commercial yield sensing system and a global navigation satellite system (GNSS) receiver. The yield data were collected for a single year in 2017 at an average sampling distance of 2 m. Each yield point was incorporated as an individual datum in all models (e.g. there was no spatial aggregation of yield data). Table 1 shows the fields, the crop harvested, descriptive yield data and the number of observations. The yield values were normalized using a Z-score to compensate for the difference in the magnitude of soy and corn yield values.

| Field | Crop   | Average yield (t/ha) | Yield standard deviation | Number of yield observations | Number of soil samples |
|-------|--------|---------------------|--------------------------|------------------------------|------------------------|
| Field A | Soybean | 4.11                | 15.4                     | 7897                         | 16                     |
| Field B | Soybean | 3.56                | 13.7                     | 14,063                       | 30                     |
| Field C | Corn   | 15.93               | 45.5                     | 20,069                       | 25                     |
| Field D | Soybean | 3.15                | 12.7                     | 3634                         | 5                      |
| Field E | Soybean | 3.46                | 11.5                     | 4612                         | 7                      |
| Field F | Soybean | 4.07                | 11.7                     | 4885                         | 11                     |
| Field G | Soybean | 3.80                | 12.1                     | 4825                         | 10                     |
| Field H | Soybean | 3.92                | 12.1                     | 5576                         | 8                      |
| Field I | Soybean | 3.78                | 30.1                     | 5158                         | 8                      |
| Field J | Soybean | 3.43                | 15.6                     | 4669                         | 7                      |
| Field K | Soybean | 3.01                | 11.9                     | 4433                         | 7                      |
| Field L | Corn   | 16.65               | 62.2                     | 10,901                       | 14                     |
| Field M | Soybean | 15.54               | 42.67                    | 11,143                       | 15                     |
| Field N | Soybean | 4.26                | 18                       | 9756                         | 13                     |
| Field O | Corn   | 14.45               | 50.7                     | 15,521                       | 13                     |
| Field P | Corn   | 14.11               | 60.5                     | 13,929                       | 17                     |
| Field Q | Soybean | 2.93                | 21.1                     | 4423                         | 19                     |

The number of soil samples taken for each field through grid sampling are also listed.
where \( x \) is the raw point minus the mean yield (\( \mu \)) divided by the yield standard deviation (\( \sigma \) of all corn or soy data pooled across all fields. Corn and soybean fields were normalized separately so that the yield of each crop type will have a mean of 0 and a standard deviation of 1, so there is a clear comparison between the different variables (Patro & Sahu, 2015).

Georeferenced grid soil samples were taken primarily in June and October of 2017 in a systematic grid so that location information would be available for each sampling point. This sampling technique provided a spatial representation of soil nutrients throughout the fields. Soil samples were taken at an average distance of 110 m apart, to a depth of approximately 150 mm. A&L Labs, a soil fertility lab accredited by the Ontario Ministry of Agriculture, Food and Rural Affairs, analyzed the samples (OMAFRA, 2017). A&L performed a chemical analysis of the samples for pH, OM, Olsen P, K, Zn and CEC using standard techniques (A&L Labs, 2017; OMAFRA, 2017).

**Topography data**

In addition to soil nutrient data, elevation data from the Southwestern Ontario Orthophotography Project (SWOOP) was also used (SWOOP, 2015). SWOOP is a 2 m horizontal resolution raster elevation product that serves as a generalized representation of both surface and ground features. While the SWOOP elevation dataset is extensively documented elsewhere, a brief description is given here. SWOOP was created using digital imagery acquired by Fugro using the Leica ADS100 geosystems sensor. The data were collected between April 12th and May 23rd, 2015, by the Ministry of Natural Resources (MRNF) in Peterborough Ontario, Canada. Imagery acquisition was undertaken at 2377 m above mean terrain (AMT) to produce a 200 mm full colour ortho-rectified image with a horizontal and vertical accuracy of 500 mm (SWOOP, 2015). While MRNF has not performed an independent error assessment of the SWOOP product, it has done so for the closely related South-Central Ontario Orthophotography Product (SCOOP). The vertical root mean squared error (RMSE) of the SCOOP product has been estimated as 186 mm (SCOOP, 2013). The SWOOP DEMs overlaying the study area were merged and required minimal data processing as both the DEM and soil samples were taken at 2 m intervals. Like the soil properties, the yield point shapefiles were overlaid on the elevation raster file, and elevation values were extracted for each crop yield point. The elevations across the 17 fields were quite similar, falling between 187 and 189 m above sea level.

The topographic wetness index was obtained for each of the segmented fields using System for Automated Geoscientific Analysis (SAGA GIS) (Conrad et al., 2015). SAGA GIS is an open-source geographic information system that is designed to implement spatial algorithms. The SAGA wetness index was applied to the elevation raster layer to reflect the theoretical distribution of lateral water accumulation (Conrad et al., 2015). The yield point shapefiles were overlaid on the SAGA wetness index raster file, and the wetness index variables were extracted for each crop yield point.

The yield data were merged with the elevation, wetness index and soil chemistry data to form a single dataset. Figure 1 provides a spatial representation of the degree of variability present in the scaled crop yield data. The blue areas represent low yield for the soybean and cornfields, while red represents high yield levels. Soil and topography spatial variability maps are available in Appendix.
Methodology

Data interpolation

To match the soil nutrient attributes to the same scale as the yield samples, this study applied three interpolation methods, Thiessen polygons, Kriging and inverse distance weighting (IDW). Thiessen polygon is polygons of influence for each sample, drawn so that each interpolated point is assigned the values to the nearest sample. All values inside the shape are equal (Panagopoulos et al., 2006). Kriging assumes that the distance or direction between sample points reflects spatial correlation used to explain variation in the surface (Chilès & Delfiner, 1999). The IDW interpolator assumed that each input point has a local influence that diminished with distance. It weights the point closer to the processing cell greater than those further away (Longman et al., 1995). Interpolation methods are a common approach for scaling in situ soil observations to field scales (for instance, Liu et al., 2015; Robinson & Metternicht, 2005; Zhang et al., 2020). However, scaling issues and field heterogeneity introduce uncertainty with interpolation methods. Thus, a semi-variogram analysis was performed to capture the spatial correlation patterns of the in-situ
origin point data to explain the predictive scaling behaviours and uncertainties of the interpolation methods (Zhang et al., 2020). Additionally, the interpolation methods were evaluated by three error metrics; the mean absolute error (MAE), root mean square (RMSE) and $R^2$. Selecting the interpolation method that produces the lowest error metrics will reduce input error in the model (Bogunovic et al., 2014). Although interpolations do not provide a definite representation of the soil characteristics, the variogram analysis and error metrics were conducted to ensure the selected interpolation methods give a fair representation of the field’s attributes.

**Variograms**

For the estimation of the extent of total sampling and analytical errors, a variographic analysis was carried out in ArcGIS software with the Geostatistical Analyst extension (Johnston et al., 2001). Since the yield samples were taken at a finer resolution than the soil samples, a semi-variogram was developed for each field and attribute to determine if the spatial distribution of the soil and topographic attributes were accommodated. The sill, nugget and range values for each attribute were then averaged to outline the ideal scale for the characteristics of interest of each field. Variogram analysis quantifies the assumption that neighboring objects appear to be more similar than those farther apart. Furthermore, it measures the strength of the statistical correlation as a distance function (Chen et al., 2015; Muukkonen et al., 2009; Zhang et al., 2020). This process may assist with future sampling processes by identifying the ideal scale at which samples should be taken for PA applications and understanding the heterogeneity of the samples. The semi-variogram consists of three key parameters: the nugget effect, sill and range. Each parameter represents a different spatial data variance characteristic. Theoretically, a variogram should go through the origin (0,0). However, when the variogram does not go through the origin, this is referred to as the nugget effect, which can be interpreted as the minimum practical error (Chen et al., 2019). The sill reflects the total variation of the spatial dataset and the partial sill, the structural variance, represents the intrinsic features of the data. The range is the distance where the variogram reaches the sill, representing the maximum spatial distance at which the dataset can still demonstrate spatial autocorrelation (Chen et al., 2019).

**Models and accuracy metrics**

For each of the interpolation soybean datasets, a MLR, ANN, decision tree and random forest model have been applied to identify which interpolation dataset was most successful for predicting yield. The models were trained with 70% of each interpolation dataset and tested with the remaining 30%. The points were split into a random train-and-test subset by the Python sklearn.model_selection.train_test_split function (Pedregosa et al., 2011). As previously mentioned, MAE, RMSE and $R^2$ were compared to evaluate each interpolation method (Supporting Information (SI), Table S-1). IDW was selected for this study as its dataset produced the lowest MAE and RMSE values for each of the models and the highest $R^2$. Although, the findings of the analysis suggests that Kriging and IDW datasets performed similarly for each of the yield models. This does not suggest that IDW is a better interpolation method than Kriging, rather the yield model performed similarly for both IDW and Kriging interpolation methods. However, the models did not perform as well
with the Theisen polygon interpolation method, which suggests that IDW and Kriging are better suited interpolation methods for this data than Theisen polygons.

Two 70/30 analyses were conducted. The first analysis focussed on the corn crop fields. The corn dataset consisted of the normalized corn yield values, and the corresponding interpolated soil nutrient and topographic values. A secondary 70/30 analysis was conducted on the separate soybean dataset and its overlying attributes to discern if the models varied in performance depending on crop type.

Three error metrics were carried out to explore the yield prediction performances: the mean absolute error (MAE), root mean square (RMSE), and $R^2$. The MAE is the average of differences in estimators (in physical units). It is represented as a percentage relative to the mean yields as yield proportions are different among crops. RMSE measures the difference between the actual and estimates, exaggerating the presence of outliers (Gonzalez-Sanchez et al., 2014; Han & Kamber, 2001). The $R^2$ value indicates how much of the variance between those two variables can be described by the linear fit. These error measurements are frequently used for agricultural systems and crop models (Jeong et al., 2016). The Gini feature importance was used for the DT, and RF models to assess which attributes had the most significant effect on the dependent variable (Pedregosa et al., 2011). Feature importance calculates each feature’s importance as the sum over the number of splits, across all trees that include the feature, proportionally to the number of samples that it splits (Altmann et al., 2010).

**Multiple linear regression**

Regression constitutes a supervised learning model, aiming to provide the prediction of how output varies according to the input variables. MLR has been a popular method for crop yield prediction (Drummond et al., 2003; Kravchenko & Bullock, 2000; Sudduth et al., 1996). The objective of MLR analysis is to study the relationship between several independent or predictor variables and a dependent or criterion variable (Adamowski et al., 2012). Equation 2 represents an MLR equation (Pedhazur, 1982):

$$Y = a + \beta_1 X_1 + \cdots + \beta_j X_j,$$

where $Y$ is a prediction of the response variable (a yield $Z$ score from either the corn or the soy dataset), $a$ is the intercept, $\beta$ is a vector of the slope or coefficients, $X$ is a vector of the predictor variables, and $j$ is the number of predictor variables. For forecasting purposes, the linear regression equation will fit a forecasting model to an observed data set of $Y$ and $X$ values. The fitted model can be used to forecast the value of $Y$ with new additional observed values of $X$ (Adamowski et al., 2012).

**Artificial neural networks**

An ANN can be used to develop empirically-based agronomic models (Kaul et al., 2005). This data-driven process captures the relationship between a large number of input and output variables from given patterns. Through these relationships, ANNs can be used to predict future values based on past histories. Commonly, the relationship obtained is non-additive and non-linear. Thus, non-linear relationships often overlooked by other techniques can be determined by ANNs with little prior knowledge of the functional relationships (Adamowski et al., 2012; Gopal & Bhargavi, 2019). In this study, the ANN model consisted of three hidden layers and used a back-propagation method to train the feed-forward
ANN. Back-propagation is often used to minimize potential errors. It is a form of supervised learning where the error rate is sent back through the network to alter the weights to improve prediction, thus decreasing error. However, a large network that uses too many nodes will become over-trained, causing the model to memorize the training data resulting in predictions with higher error. The process is repeated until either a specified error limit is achieved or the total number of training cycles (epochs) has been completed (Gonzalez-Sanchez et al., 2014; Kaul et al., 2005; Khairunniza-Bejo et al., 2014). Keras, a deep learning application programming interface (API) written in Python, was applied to develop the ANN model (Chollet, 2015). Models in Keras are defined as a sequence of layers, so first a sequential model is generated, and three additional layers were added utilizing the ReLu activation function. The number of additional layers was selected through trial and error. Fewer or more hidden layers resulted in greater MAE and RMSE values and a lower R². Once the model is defined, the model is then compiled with a mean squared error loss function and Adam optimizer. The loss function repeatedly estimates the error so that the weights can be updated to reduce the loss on the next evaluation, while the Adam optimizer improves the speed of convergence and finds a better minimum for the loss function (Yi et al., 2020). The model is then trained or fit on the training dataset before predicting the test dataset.

**Decision trees**

A decision tree (DT) consists of a flow-chart-like tree structure where various aspects and attributes are considered for evaluating an issue. A recursive algorithm was used for further assessment of classifying the attribute with the highest information (Elavarasan et al., 2018). Furthermore, when a DT is used for prediction, it is assumed that the response variable’s nature is continuous (Raorane & Kulkarni, 2012; Veenadhari et al., 2011). A DT model is built from data or observations according to some criteria. The model aims to learn a general rule from the observed instances (Raorane & Kulkarni, 2012; Veenadhari et al., 2011). The dataset is gradually organized into small homogenous subsets, while a corresponding tree graph is generated (Liakos et al., 2018). The first node in the tree is named the root node (Gonzalez-Sanchez et al., 2014). Each internal node of the tree structure represents a different pairwise comparison on a selected feature, whereas each branch represents the outcome of this comparison (Liakos et al., 2018). A node with outgoing edges is referred to as a test node, whereas a node without outgoing edges is called a leaf node (Gonzalez-Sanchez et al., 2014). Leaf nodes represent the final decision or prediction after following the root-to-leaf path (expressed as a rule of classification). The variance reduction algorithm was used for continuous target variables (Liakos et al., 2018). This algorithm used the standard formula of variance to choose the best split. In this study, the scikit-learn Python module “DecisionTreeRegressor” class was applied with a max depth of 30 trees, which was selected based on the R² value (Pedregosa et al., 2011). After 30 trees, the model becomes over-trained, resulting in similar or weaker predictions. A variance reduction algorithm was used to choose the best split (Liakos et al., 2018). The split with lower variance was selected as the criteria to split the population.

**Random forest**

Random forest (RF) is a non-parametric advanced classification and regression tree (CART) analysis method that consists of multiple decision trees (Jeong et al., 2016).
Each tree is built from a random sample of the training data and is drawn with replacements. The data is recursively split into more homogenous units, referred to as nodes, to improve the response variable’s predictability. The most efficient split is defined by identifying the predictor variable and the split point that results in the largest reduction in the residual sum of squares between the sample observations and the node mean. All trees are grown to the maximum extent that is controlled by the size of the nodes. The result is an ensemble of low bias and high variance regression trees, where the final predictions are derived by averaging the predictions of the individual trees (Aghighi et al., 2018; Jeong et al., 2016; Kern et al., 2019).

Similar to the DT models, the RF models were built using the scikit-learn Python module “RandomForestRegressor” class with 30 trees derived from bootstrapped data-sets (Pedregosa et al., 2011). Bootstrap aggregation attempts to mitigate the problems of high variance and high bias of the final prediction model through a reduction of the correlation between estimators. The minimum number of samples required to be tested at each node was set to the default value of the square root of the total number of predictor variables used. Additionally, the maximum depth of the tree was set to 30, to aid in the comparison between the RF and DT models (Gopal & Bhargavi, 2019; Jeong et al., 2016).

**Cross-validation techniques**

Three different cross-validation methods were used to evaluate the models further. First, a “Jack-Knifing” approach was applied to each field. This approach consisted of eliminating all of the yield points for one of the fields and predicting the missing yield values with the remaining yield estimates. The “Jack-Knifing” technique used both the DT and RF models to identify potential outlier fields. Additionally, fields that had a high error in this analysis were identified as outliers as the surrounding fields were not compatible enough to provide an accurate prediction. For each field, a feature importance calculation was taken to determine the independent variables that had the most significant impact on yield.

Next, this study used a “Leave-Group-Out” RF cross-validation method. This method was similar to the "Jack-Knifing" approach, although groups of three fields were selected instead of singular fields. The yield data of these three fields were removed, and the training dataset comprised of the remaining individual fourteen fields datasets. The groups were selected based on the results from the “Jack-Knifing” approach and consisted of five separate trials containing different missing fields. Trial 1 consisted of the fields that had the lowest $R^2$ value and highest error in the "Jack-Knifing" analysis, while Trial 2 comprised of the top-performing fields. Trial 3 included three fields that had $R^2$ values around the median. Finally, Trial 4 and 5 were similar in which fields that varied in performance were selected.

Finally, a model complexity reduction method was applied to the RF model. This process involved eliminating one attribute at a time based on feature importance from the 70/30 training and testing analysis. The RF model was re-trained each time an attribute had been removed and then compared with $R^2$ and error metrics. The first method of reduction removed attributes with the lowest feature importance values first. Alternatively, the second model of reduction removed the attributes with the highest feature importance values first. The model reduction technique provided insight into
which attributes were necessary for yield prediction and was used to cross-validate the feature importance analysis.

**Results**

**Spatial structure analysis**

In geostatistical theory, the range of the semi-variogram is the maximum distance of the correlated measurements. It can be a sufficient criterion for the selection of sampling design in mapping soil properties (Utset et al., 1998). As a rough guide, the sampling interval should be less than half the variogram range (Kerry & Oliver, 2004; Mallarino et al., 2006). Based on the range of the variograms for CEC, Zn and yield, the maximum sampling interval that exhibits correlation between measurements to support the interpolation estimates approximately 300 m. Potassium could achieve a suitable sampling interval at around 375 m, while OM at 500 m, elevation and wetness index at 200 m, and finally pH and soil test phosphorus (STP) at 130 m. While pH and STP may have scales not resolved by the sampling density in this study, it is unlikely that the other attributes do. The response variable (yield) also exhibited a range of 610 m, suggesting that finer scale variations of, e.g., STP and pH do not influence yield significantly. These results suggest that the sampling intervals of 110 m for this study were sufficient for modeling.

**Correlation analysis**

A Pearson correlation matrix summarized the relationship between the soil and topography data of the scaled corn and soybean datasets (Tables 2 and 3). The correlation matrix showed a positive correlation between soil nutrients. However, the correlation matrix did not provide a clear interpretation regarding soil nutrients and topography concerning yield suggesting the relationships may be non-linear.
Model comparison

As previously mentioned, performance metrics evaluated the outcomes of the 70/30 comparison of MLR and ML models. Table 4 shows the results for the MAE, RMSE and $R^2$ metrics of the corn dataset and soybean dataset in the analysis. The models performed similarly, with DT and RF models performing with the lowest error and highest $R^2$. The soybean dataset achieved a higher $R^2$ and lower error metrics than corn. However, this is likely due to the soybean dataset being larger than the corn dataset.

Figure 2 shows the feature importance of each variable for the RF and DT models. In this study, both DT and RF models’ feature importance shows that STP and pH have the highest values in all datasets. These results from the 70/30 analysis suggest that STP and pH are more significant than the other features and lead to bigger information gains within the models. As well, the similar feature importance values indicate that the corn and soybean models were dependent on the same variables, and the merged dataset is sufficient to represent both crops.

“Jack-Knifing” cross-validation

The models of the best performance from the 70/30 training and testing analysis, DT and RF, were utilized in the “Jack-Knifing” cross-validation technique. The “Jack-Knifing” analysis include both corn and soybean fields to account for the data needs of the ML models and increase the variability in the training set. Additionally, corn and soybean field identifiers were included in the merged dataset to distinguish crop types. This analysis consisted of erasing one field yield values and then applying the remaining sixteen fields’ datasets to predict the missing yield values. Random forest and DT models were both used in this analysis as they had the lowest mean error for both metrics and the highest $R^2$ values in the 70/30 analysis. These results suggest that both RF and DTs are reliable for quantifying the relationship between crop yield, soil properties and topographic characteristics for predicting crop yield in this case. Figure 2 shows the “Jack-Knifing” results for each field.

However, there were three fields that the “Jack-Knifing” approach identified as outliers for both DT and RF models. Field A, Field B and Field C had the highest mean error and the lowest $R^2$ value. This study did not include field identification attributes in the model analyses, so Fields A, B and C were anonymized. Additionally, Fields A, B and C are contiguous, as presented in Fig. 3.

| Yield | Elevation | Wetness | pH | K | OM | CEC | STP | Zn |
|-------|-----------|---------|----|---|----|-----|------|----|
| Yield | 1.00      | -0.32   | -0.07 | -0.08 | -0.56 | -0.06 | 0.38  | -0.49 | -0.30 |
| Elevation | -0.32 | 1.00   | -0.25 | -0.42 | 0.26 | -0.07 | -0.58 | 0.29 | 0.54 |
| Wetness | -0.07 | -0.25 | 1.00 | 0.07 | 0.14 | 0.24 | 0.05 | 0.17 | 0.14 |
| pH | -0.08 | -0.42 | 0.07 | 1.00 | 0.23 | 0.25 | 0.24 | 0.20 | -0.23 |
| K | -0.56 | 0.26 | 0.14 | 0.23 | 1.00 | 0.46 | -0.03 | 0.81 | 0.60 |
| OM | -0.06 | -0.07 | 0.24 | 0.25 | 0.46 | 1.00 | 0.13 | 0.36 | 0.32 |
| CEC | 0.38 | -0.58 | 0.05 | 0.24 | -0.03 | 0.13 | 1.00 | -0.29 | -0.32 |
| STP | -0.49 | 0.29 | 0.17 | 0.20 | 0.81 | 0.36 | -0.29 | 1.00 | 0.77 |
| Zn | -0.30 | 0.54 | 0.14 | -0.23 | 0.60 | 0.32 | -0.32 | 0.77 | 1.00 |
Through the feature importance analysis, STP and pH were selected as the most significant attributes in relevance to yield for the majority of fields. These results further support the feature importance analysis conducted in the 70/30 model results. STP and pH had the highest values in the feature importance analysis for Fields A, B and C in the "Jack-Knifing" examination. However, due to the model’s high error, these attributes are not relevant to yield for these fields. Furthermore, Field M feature importance analysis identified CEC and K as the most relevant variables to yield for both DT and RF models. This difference in feature importance indicates that Field M would potentially need a tailored soil management strategy to compensate for the variable importance discrepancy compared with other fields.

"Leave-Group-Out" cross-validation

The cross-validation technique "Leave-Group-Out" was used to determine how successful the models were in predicting a large amount of missing data. Like the "Jack-Knifing" approach, the "Leave-Group-Out" analysis consisted of erasing the yield data of three

| Table 4 | Mean absolute error (MAE), root mean square error (RMSE) and $R^2$ metrics results for all the evaluated techniques in the 70/30 analysis |
|---------|---------------------------------------------------------------------------------|
| Method  | MAE  | RMSE  | $R^2$  |
|---------|------|-------|--------|
| Corn    |      |       |        |
| MLR     | 40.5 | 61.8  | 0.40   |
| ANN     | 33.1 | 50.6  | 0.68   |
| DT      | 12.9 | 35.4  | 0.80   |
| RF      | 12.0 | 31.3  | 0.85   |
| Soybean |      |       |        |
| MLR     | 45.9 | 58.8  | 0.45   |
| ANN     | 21.9 | 32.2  | 0.83   |
| DT      | 8.8  | 23.9  | 0.91   |
| RF      | 7.5  | 20.0  | 0.94   |

Fig. 2 Feature importance results for both the decision tree (DT) and random forest (RF) models for all the independent variables in relation to yield in the 70/30 training model
fields, while the training dataset consisted of the remaining fourteen fields datasets. The groups were selected based on the results from the “Jack-Knifing” approach and consisted of five separate trials comprising of different missing fields. With the missing yield data, the RF model predicted approximately 16% missing data for the “Leave-Group-Out” cross-validation method. The RF model was selected for this method as it had slightly outperformed the DT model in the “Jack-Knifing” examination. Trial 1 consisted of eliminating the yield values for three fields with the highest error in the “Jack-Knifing” analysis. Trial 1 had the highest RMSE value, and the model explained 6.5% of the yield variation. The poor performance of Trial 1 suggests that there are factors not accounted for in this study that control yield in those fields. As such, fields in Trial 1 likely cannot be managed the same way as the other fields. Different fields or attributes may improve the model’s performance, although that is outside the scope of this study.

Trial 2 removed the yield values of the three best-performing fields from the “Jack-Knifing” approach. The predicted yields matched the observed data well with the model

---

**Table 5** Mean absolute error (MAE), root mean square error (RMSE) and R² metrics results for all the evaluated trials for the “Leave-Group-Out” cross-validation analysis. The identity of the fields that had missing yield values for each trial was also included.

| Missing Fields (to be Predicted from model trained at the other fields) | MAE  | RMSE | R²  |
|-----------------------------------------------------------------------|------|------|-----|
| Trial 1: A, B, C                                                      | 43.4 | 74.7 | 0.06|
| Trial 2: D, E, M                                                      | 11.1 | 23.6 | 0.92|
| Trial 3: J, E, G                                                      | 11.2 | 31.2 | 0.87|
| Trial 4: I, B, D                                                      | 20.1 | 39.5 | 0.76|
| Trial 5: Q, I, K                                                      | 21.5 | 37.2 | 0.78|

---

**Fig. 3** The locations and shape of the 17 fields located in Southwestern, Ontario.
explaining 92% of the yield variation. As well, Trial 2 had the lowest RMSE of the five trials. Trial 3 removed the yield values of three fields of average performance and explained 87% of the yield variation. Trial 4 and 5, which had fields that ranged in performance, had acceptable R² values and moderate error values, as shown in Table 5. These cross-validation results demonstrate that while soil chemistry-yield relationships can often be generalized to other fields not used in calibration, there are some cases where they cannot be. Future research is needed to establish which fields can be expected to have similar soil chemistry-yield relationships. Additionally, the “Leave-Group-Out” cross-validation demonstrates that some groups of fields share a similar relationship of soil to yield, as shown through the satisfactory performance of Trials 2 and 3. Due to their poor performance in Trial 1, it is clear that the fields A, B and C do not share the same soil-to-yield relationship as the remaining fields. In future studies, an additional analysis including other properties, such as climatic variables or alternative fields, may assist in identifying how fields in Trial 1 differ.

Model reduction cross-validation

The attribute reduction method assisted in determining the most significant predictors. The ranking of the DT and RF feature importance values guided the attribute reduction analysis. The RF model was utilized in this method as it had slightly outperformed the DT in the previous cross-validation methods. Initially, all the attributes were included in the model and had an R² value of 0.93 (Fig. 4). The wetness index was removed from the dataset as it had the lowest feature importance value, and the RF model was a rerun. With the wetness index removed, the R² value only decreased by 0.004. Furthermore, elevation, K, OM and Zn were removed from the dataset in sequence; however, with the removal of the following attributes, the R² only dropped to 0.91 with relatively low error. These results suggest that CEC, pH and STP account for 91% of the yield variation. Even with only pH and STP attributes, which had the highest feature importance values, the model maintained an R² of 0.56.
This study conducted a second attribute reduction analysis; however, this time, the variables with the highest feature importance values were deducted first (Fig. 5). For instance, STP was the first attribute to be removed, and the $R^2$ value was reduced by 0.02. However, with the removal of STP, pH and CEC, the model’s $R^2$ dropped to 0.88. Although this is still an acceptable $R^2$, the rate at which the $R^2$ value dropped and the increase in error metrics is more significant compared to the previous model reduction experiment. When K, elevation and moisture were the only attributes remaining, the model’s $R^2$ value decreased to 0.73.

**Discussion**

This study focused on comparing regression and ML techniques to assess the relationships between soil and topographic properties to assist with crop yield predictions. The RF model was the most successful in the 70/30 comparison with an $R^2$ value greater than 0.85 for both soybean and corn models, and MLR models were the least successful. In addition to evaluating the predictive capabilities of the models, additional cross-validation techniques were applied to evaluate the relationships between fields and to determine which attributes were most relevant to crop yield prediction. The Gini feature importance suggests that STP and pH were the most significant attributes when predicting crop yield in the 70/30 analysis. However, the attribute reduction analysis suggests that the importance of attributes may vary. The “Jack-Knifing” and “Leave-Group-Out” cross-validation techniques show that RF and DTs can predict low- and high-yield areas for different fields if they are similar.

**Incorporating soils information into precision agriculture**

Several studies have already compared the predictive potential of regression and ML models with primarily climatic and farm management attributes, as described above (Gonzalez-Sanchez et al., 2014). However, few studies have explored the ML model’s abilities to regress crop yields with only soil nutrient and topography data in Southwestern Ontario (Jeong et al., 2016). The results from this study illustrate that the RF and DT models were
sufficient at predicting crop yield with only soil and topography attributes and they outperformed MLR and ANN in the 70/30 testing analysis.

Many variables related to crop production are often strongly correlated with and within each other. Through the correlation matrix, it was evident that there is multicollinearity present among the attributes. For instance, for both corn and soy, K and STP share a strong positive correlation likely due to co-application in fertilizer. Random forest and DT models have an advantage when predictor or explanatory variables are highly correlated. These models use the single best variable when splitting responses at each node and average the predictions of the trees in the forest to make a multidimensional step function. This process suggests that even if multiple variables are correlated and similarly drive the response, only one can affect the RF and DT model at a time (Jeong et al., 2016). However, variable collinearity can be a critical issue in conventional regression models. A limitation in regression models is the presumption that there is little to no collinearity between variables and independent and dependent relationships are linear (Meersmans et al., 2008).

The feature importance was used to identify the most important soil attributes for crop yield variability. A cursory review of the results might suggest that STP and pH are the most influential variables in the 70/30 training and testing model (Table 5). The field-by-field results presented in Fig. 6 show that while there is some variability in variable importance in space, the overall pattern is the same as for the pooled dataset shown in Table 4. The model reduction experiment shown in Fig. 4 is also consistent with this view, as withholding all the variables from the model except for STP and pH (depicted in the CEC bar) resulted in a model fit nearly as high as that obtained from using all of the soil variables. However, the results of the model reduction experiment shown in Fig. 5 challenges this view, as it is possible to have a very high model fit without STP or pH. The correlation matrices of soil variables presented in Tables 3 and 4 show that yield tends to be higher in areas with lower elevation, lower nutrient contents of K, STP and Zn, and higher contents of OM and CEC. These patterns are consistent with recent work by Pluer et al. (2020), who sampled a single field in Southern Ontario intensively for soil attributes and used NDVI to estimate yield. They found similar patterns, with yield estimates being highest in areas of lowest nutrient content and highest OM and CEC surrogates (clay content). The results suggest that the RF and DT models are identifying this pattern, not that STP or pH are, by themselves, strong drivers of crop variability. Indeed, the negative association of STP and yield suggests that yield is hindered by other drivers and is unable to take up all the phosphorus applied as fertilizer. These results point to the importance of OM in crop

![Fig. 6](attachment:image-url) Field identification (ID) and $R^2$ metrics results for the decision tree (DT) and random forest (RF) analyses in the “Jack-Knifing” analysis.
productivity, and also to the importance of crop management that is tailored to the unique conditions within fields.

It is well established that the continued application of fertilizers and manures in many agricultural areas has led to accumulations of soil phosphorus in excess of local crop needs (Sharpley et al., 2001). While phosphorus is an essential nutrient for plant growth, runoff of phosphorus can accelerate eutrophication resulting in severe impairment of water bodies’ abilities to support aquatic, recreational and drinking water uses (Sharpley et al., 2001). STP is positively correlated with dissolved phosphorus in runoff from soils in Southern Ontario (Wang et al., 2010, 2015), leading Mohamed et al. (2019) to suggest that addressing the re-eutrophication of Lake Erie may require reductions to STP. This work and others’ show that within fields in Southwest Ontario, STP is negatively correlated with yield. This suggests there may be ways to reduce STP such that yield is unaffected or possibly increases. Future work with ML and soil properties could try to identify areas where high STP is not contributing to increased yield. Indeed, Liu et al. (2015) highlighted that a 10-year fertilizer cessation had not affected wheat yields in Saskatchewan, Canada, although STP had decreased.

As previously mentioned, Southwestern Ontario, is relatively flat. The 17 fields predominately had a slope of 0°, with a maximum slope of 13°. Several studies have suggested that topographic attributes have a significant correlation with yield. For instance, Changere and Lal (1997), Kravchenko and Bullock (2000), and McConkey et al. (1997) observed higher yields at lower slope positions and lower yields at high positions. However, in this study, elevation and the topographic wetness index were not given high feature importance in predicting yield, as shown in Table 4. There is the potential that the elevation data were not resolved finely enough to detect significant differences in height and slope. It is likely, the low slope values in the fields are accounting for the lack of correlation between topography and yield. Due to the low slope, there is potentially little water redistribution, little erosion and limited soil redistribution.

Generalizing the relationship between soil attributes and yield

The “Jack-Knifing” and "Leave-Group-Out" cross-validation methods suggest that fields of similar soil and topographic characteristics may assist in predicting missing yield. This analysis can be useful for predicting yield for a field that has not yet been sampled or has missing data. For instance, with 16% missing data in the “Leave-Group-Out” analysis, the RF model was successful at predicting yield at a relatively high level of accuracy. However, for the model to perform well, the fields need to have similar topologic and soil characteristics as the surrounding fields. For instance, in Trial 1 of the “Leave-Group-Out” analysis, Fields A, B and C were selected, the model performed poorly with high error and a low $R^2$. As well, these fields had performed poorly in the "Jack-Knifing" cross-validation assessment. The inability to predict yield for Fields A, B and C are likely dependent on attributes that were not considered in this study, such as farm management practices or additional soil attributes. By identifying fields where the relationship between soil and yields is fundamentally different from others, unique management areas could be delineated.

There is some evidence in the study that grid sampling alone could be used to anticipate yield maps, as the relationships between in-field soil properties and yield often held between different fields. However, doing so would require knowing in advance which fields had similar soil property-yield relationships. By splitting the fields into similar classes, users can apply the understanding of one group of fields to another. Future research could focus on ways to anticipate which soil property-yield relationships would hold.
Conclusion

This study evaluated the effectiveness of MLR and ML techniques in modeling complex yield responses of corn and soybean crops in 17 fields. In the 70/30 analysis, RF and DT models outperformed the other ML and regression algorithms in predicting yield. The cross-validation techniques took the comparative analysis a step further. They identified which attributes had a significant impact on yield, predicted yield with up to 16% missing data, and suggested the potential for PA upscaling with limited data if field characteristics are similar. These analyses have shown that soil nutrients are useful in describing yield variability on an agricultural field scale. The soil characteristics were particularly useful in site-specific management for delineating areas. Topographic properties did not play a significant role in predicting crop yield. Nevertheless, topographic data in fields with more significant variation in the elevation and wetness index are likely to have a more notable effect on yield.

The model reduction analysis demonstrated that although STP and pH had the highest feature importance analysis for most examinations, the correlation among soil properties complicated the interpretation. For instance, K and OM were also significant in the crop yield prediction models. Additionally, cross-validation methods were effective in identifying outlier fields. Additional attributes or fields are likely necessary to increase the diversity of the dataset to improve the predictive ability of the models in fields identified as outliers.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11119-022-09897-0.

Author contributions CW obtained funding for the research; CW and HB designed the research; HB conducted the research, including coding and data management, with input and guidance from Christopher Wellen; and Hannah Burdett wrote the paper, with input and guidance from CW.

Funding NSERC Discovery Grant to CW, Canada-Ontario Agreement Grants COA/GLS 1009C and 1010D-16.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The author(s) declare(s) that they have no competing interests.

References

Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B., & Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resources Research. https://doi.org/10.1029/2010WR009945
Aghighi, H., Azadbakht, M., Ashourloo, D., Shahrabi, H. S., & Radiom, S. (2018). Machine learning regression techniques for the silage maize yield prediction using time-series images of landsat 8 OLI. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(12), 4563–4577.
Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. Bioinformatics, 26(10), 1340–1347.
Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Chollet, F. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras

Chilès, J. P., & Delfiner, P. (1999). Wiley-Interscience. Geostatistics: Modeling spatial uncertainty

Corwin, D. L., & Lesch, S. M. (2003). Application of soil electrical conductivity to precision agriculture. Agronomy Journal, 95(3), 455–471.

Dahikar, S. S., & Rode, D. S. V. (2014). Agricultural Crop Yield Prediction Using Artificial Neural Network Approach. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, 2(1), 683–686.

D’Amario, S. C., Rearick, D. C., Fasching, C., Kemble, S. W., Porter-Goff, E., Spooner, D. E., Williams, C. J., Wilson, H. F., & Xenopoulos, M. A. (2019). The prevalence of nonlinearity and detection of ecological breakpoints across a land use gradient in streams. Scientific Reports, 9(1), 1–11.

Drummond, S. T., Sudduth, K. A., Joshi, A., Birrell, S. J., & Kitchen, N. R. (2003). Statistical and neural methods for site–specific yield prediction. Transactions of the ASAE, 46(1), 5.

Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y., & Srinivasan, K. (2018). Forecasting yield by integrating agrarian factors and machine learning models: A survey. Computers and Electronics in Agriculture, 155, 257–282.

Frank, R., & Ripley, B. D. (1977). Land use activities in eleven agricultural watersheds in Southern Ontario, Canada. Ontario Ministry of Agriculture and Food, International Reference Group on Great Lakes Pollution from Land Use Activities, 1975–1976. International Joint Commission (IJC) Digital Archive. https://scholar.uwindsor.ca/ijchaparchive/122

Gopal, P. M., & Bhardwaj, R. (2019). A novel approach for efficient crop yield prediction. Computers and Electronics in Agriculture, 165, 104968.

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. Spanish Journal of Agricultural Research, 12(2), 313–328.

Han, J., & Kamber, M. (2001). Data mining concepts and techniques (pp. 335–391). Morgan Kaufmann Publishers.

International Society of Precision Agriculture. (2019). Precision agriculture definition: International Society of Precision Agriculture. Retrieved October 29, 2021, from https://www.ispag.org/about/definition

Jeong, J.-H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R., & Kim, S.-H. (2016). Random forests for global and regional crop yield predictions. PLoS ONE. https://doi.org/10.1371/journal.pone.0156571

Johnston, K., Ver Hoef, J. M., Krivoruchko, K., & Lucas, N. (2001). Using ArcGIS geostatistical analysis. Environmental Systems Research Institute.

Jung, W. K., Kitchen, N. R., Sudduth, K. A., & Anderson, S. H. (2006). Spatial characteristics of claypan soil properties in an agricultural field. Soil Science Society of America Journal, 70(4), 1387–1397.

Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. Agricultural Systems, 85(1), 1–18.

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. Survey Research Methods, 13(1), 73.

Kerr, J. M., DePinto, J. V., McGrath, D., Sowa, S. P., & Swinton, S. M. (2016). Sustainable management of Great Lakes watersheds dominated by agricultural land use. Journal of Great Lakes Research, 42(6), 1252–1259.

Kerry, R., & Oliver, M. A. (2004). Average variograms to guide soil sampling. International Journal of Applied Earth Observation and Geoinformation, 5(4), 307–325.

Khairunniza-Bejo, S., Mustaffa, S., & Ismail, W. I. W. (2014). Application of artificial neural network in predicting crop yield: A review. Journal of Food Science and Engineering, 4(1), 1.
Khazaei, J., Naghavi, M. R., Jahansouz, M. R., & Salimi-Khorshidi, G. (2008). Yield estimation and clustering of chickpea genotypes using soft computing techniques. *Agronomy Journal, 100*(4), 1077–1087.

Kravchenko, A. N., & Bullock, D. G. (2000). Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy Journal, 92*(1), 75–83.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochitis, D. (2018). Machine learning in agriculture: A review. *Sensors, 18*(8), 2674.

Liu, J., Goering, C. E., & Tian, L. (2001). A neural network for setting target corn yields. *Transactions of the ASAE, 44*(3), 705.

Liu, J., Hu, Y., Yang, J., Abdi, D., & Cade-Menun, B. J. (2015). Investigation of soil legacy phosphorus transformation in long-term agricultural fields using sequential fractionation, P K-edge XANES and solution P NMR spectroscopy. *Environmental Science & Technology, 49*(1), 168–176.

Longman, R. H. G., Ter Braak, C. J. F., & Van Tongeren, O. F. R. (1995). *Data analysis in community and landscape ecology*. Cambridge University Press.

Mallarino, A. P., Beegle, D. B., & Joern, B. C. (2006). Soil sampling methods for phosphorus-spatial concerns. Southern Education Research Activities (SERA) 17, United States Department of Agriculture.

McConkey, B. G., Ulrich, D. J., & Dyck, F. B. (1997). Slope position and subsoiling effects on soil water and spring wheat yield. *Canadian Journal of Soil Science, 77*(1), 83–90.

Meersmans, J., De Ridder, F., Canters, F., De Baets, S., & Van Molle, M. (2008). A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). *Geoderma, 143*(1–2), 1–13.

Metwally, M. S., Shaddad, S. M., Liu, M., Yao, R. J., Abdo, A. I., Li, P., Jiao, J., & Chen, X. (2019). Soil properties spatial variability and delineation of site-specific management zones based on soil fertility using fuzzy clustering in a hilly field in Jianyang, Sichuan, China. *Sustainability (switzerland), 11*(24), 7084.

Miao, Y., Mulla, D. J., & Robert, P. C. (2006). Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. *Precision Agriculture, 7*(2), 117–135.

Mittal, G. S., & Zhang, J. (2000). Prediction of temperature and moisture content of frankfurters during thermal processing using neural network. *Meat Science, 55*(1), 13–24.

Mohamed, M. N., Wellen, C., Parsons, C. T., Taylor, W. D., Arhonditis, G., Chomicki, K. M., Boyd, D., Weidman, P., Mundle, S. O. C., Van Cappellen, P., Sharples, A. N., & Haffner, D. G. (2019). Understanding and managing the re-eutrophication of Lake Erie: Knowledge gaps and research priorities. *Freshwater Science, 38*(4), 675–691.

Muukkonen, P., Häkkinnen, M., & Mäkipää, R. (2009). Spatial variation in soil carbon in the organic layer of managed boreal forest soil—implications for sampling design. *Environmental Monitoring and Assessment, 158*(1), 67–76.

Mzuku, M., Khosla, R., Reich, R., Inman, D., Smith, F., & MacDonald, L. (2005). Spatial variability of measured soil properties across site-specific management zones. *Soil Science Society of America Journal, 69*(5), 1572–1579.

Nelligan, C., Sorichetti, R. J., Yousif, M., Thomas, J. L., Wellen, C. C., Parsons, C. T., & Mohamed, M. N. (2021). Then and now: Revisiting nutrient export in agricultural watersheds within southern Ontario’s lower Great Lakes basin. *Journal of Great Lakes Research, 47*(6), 1689–1701.

[OMAFRA] Ontario Ministry of Agriculture, Food and Rural Affairs. (2017). Agronomy Guide for Field Crops. Publication 811. Ontario Ministry of Agriculture, Food and Rural Affairs.

Panagopoulos, T., Jesus, J., Antunes, M. D. C., & Beltrao, J. (2006). Analysis of spatial interpolation for optimising management of a salinized field cultivated with lettuce. *European Journal of Agronomy, 24*(1), 1–10.

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture, 121*, 57–65.

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. Non-peer reviewed preprint at arXiv preprint. arXiv:1503.06462.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Prediction and explanation*. Holt, Rinehart, & Winston.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.

Pluer, E. M., Robinson, D. T., Meinen, B. U., & Macrae, M. L. (2020). Pairing soil sampling with very-high resolution UAV imagery: An examination of drivers of soil and nutrient movement and agricultural productivity in southern Ontario. *Geoderma, 379*, 114630.
Raorane, A. A., & Kulkarni, R. V. (2012). Data Mining: An effective tool for yield estimation in the agricultural sector. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* 1(2), 1–4.

Richards, N. R., Caldwell, A. G., & Morwick, F. F. (1949). Soil survey of Essex County (No. 11). Experimental Farms Service, Dominion Department of Agriculture and the Ontario Agricultural College.

Robertson, M., Isbister, B., Maling, I., Oliver, Y., Wong, M., Adams, M., Bowden, B., & Tozer, P. (2007). Opportunities and constraints for managing within-field spatial variability in Western Australian grain production. *Field Crops Research,* 104(1–3), 60–67.

Robinson, T. P., & Metternicht, G. (2005). Comparing the performance of techniques to improve the quality of yield maps. *Agricultural Systems,* 85(1), 19–41.

SCOOP. (2013). Land Information Ontario Imagery, South-Central Ontario Orthophotography Product, Ontario Ministry of Natural Resources and Forestry. Retrieved February, 2022, from https://geohub.lio.gov.on.ca/documents/442deae4b894470a57821a2b48f783e/about

Seyhan, A. T., Tayfur, G., Karakurt, M., & Tanoglu, M. (2005). Artificial neural network (ANN) prediction of compressive strength of VARTM processed polymer composites. *Computational Materials Science,* 34(1), 99–105.

Sharpley, A. N., McDowell, R. W., & Kleinman, P. J. (2001). Phosphorus loss from land to water: Integrating agricultural and environmental management. *Plant and Soil,* 237(2), 287–307.

Sudduth, K. A., Drummond, S. T., Birrell, S. J., & Kitchen, N. R. (1996). Analysis of spatial factors influencing crop yield. In P. C. Robert, R. H. Rust, & W. E. Larson (Eds.), *Proceedings of the third international conference on precision agriculture* (pp. 129–139). Crop Science Society of America.

SWOOP. (2015). Ontario Digital Elevation Model. Land Information Ontario, Ontario Ministry of Natural Resources and Forestry. Retrieved October, 2018, from https://geohub.lio.gov.on.ca/maps/mnr::ontario-digital-elevation-model-imagery-derived/about

Tan, C. S., & Reynolds, W. D. (2003). Impacts of recent climate trends on agriculture in southwestern Ontario. *Canadian Water Resources Journal,* 28(1), 87–97.

Tantalaki, N., Souravlas, S., & Roumeliotis, M. (2019). Data-driven decision making in precision agriculture: The rise of big data in agricultural systems. *Journal of Agricultural & Food Information,* 20(4), 344–380.

Tey, Y. S., & Brindal, M. (2012). Factors influencing the adoption of precision agricultural technologies: A review for policy implications. *Precision Agriculture,* 13(6), 713–730.

Utset, A., Ruiz, M. E., Herrera, J., & de Leon, D. P. (1998). A geostatistical method for soil salinity sample site spacing. *Geoderma,* 86(1–2), 143–151.

Wang, Y. T., Zhang, T. Q., Hu, Q. C., Tan, C. S., Halloran, I. O., Drury, C. F., Reid, D. K., Ma, B. L., Ball-Coelho, B., Lauzon, J. D., Reynolds, W. D., & Welacky, T. (2010). Estimating dissolved reactive phosphorus concentration in surface runoff water from major Ontario soils. *Journal of Environmental Quality,* 39(5), 1771–1781.

Wang, Y. T., Zhang, T. Q., O’Halloran, I. P., Hu, Q. C., Tan, C. S., Speranzini, D., Macdonald, L., & Patterson, G. (2015). Agronomic and environmental soil phosphorus tests for predicting potential phosphorus loss from Ontario soils. *Geoderma,* 241, 51–58.

Veenadhari, S., Misra, B., & Singh, C. D. (2011). Data mining techniques for predicting crop productivity—A review article. *International Journal of Computer Science and Technology,* 2(1), 98–100.

Vollmer-Sanders, C., Allman, A., Busdeker, D., Moody, L. B., & Stanley, W. G. (2016). Building partnerships to scale up conservation: 4R Nutrient Stewardship Certification Program in the Lake Erie watershed. *Journal of Great Lakes Research,* 42(6), 1395–1402.

Yi, D., Ahn, J., & Ji, S. (2020). An effective optimization method for machine learning based on ADAM. *Applied Sciences,* 10(3), 1073.

Zhang, Y., Zhen, Q., Li, P., Cui, Y., Xin, J., Yuan, Y., et al. (2020). Storage of soil organic carbon and its spatial variability in an agro-pastoral ecotone of Northern China. *Sustainability,* 12(6), 2259.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.