The International Society of Precision Agriculture presents the
16th International Conference on
Precision Agriculture
21–24 July 2024 | Manhattan, Kansas USA

Predicting soil chemical properties using proximal soil sensing technologies and topography data: A case study

Felippe H. S. Karp[1], Viacheslav Adamchuk[1], Pierre Dutilleul[2], Alexei Melnitchouck[3], Asim Biswas[4]

[1]Bioresource Engineering Department, McGill University, Canada; [2]Department of Plant Science, McGill University, Canada; [3]Intelmax Corp., Canada; [4]School of Environmental Sciences, University of Guelph

A paper from the Proceedings of the
16th International Conference on Precision Agriculture
21-24 July 2024
Manhattan, Kansas, United States

Abstract.

Using proximal soil sensors (PSS) is widely recognized as a strategy to improve the quality of agricultural soil maps. Nevertheless, the signals captured by PSS are complex and usually relate to a combination of processes in the soil. Consequently, there is a need to explore further the interactions at the source of the information provided by PSS. The objectives of this study were to examine the relationship between proximal sensing techniques and soil properties and evaluate the feasibility of using data fusion to improve the mapping of soil chemical properties with extra-low sampling densities. Field data from ground penetrating radar, passive gamma-ray spectrometry, apparent electrical conductivity, resistance to penetration, and elevation were collected from a 43-ha site in Central Alberta, Canada. Soil sampling (originally with 0.4 ha·sample^{-1} density) and subsequent lab analysis provided information on soil organic matter, pH, and plant-available phosphorous (P) and potassium (K). After pre-processing and co-locating the sampling and sensor data, soil properties and some PSS data were correlated. Samples were then removed until a density of 3.5 ha·sample^{-1} was reached, thus creating an extra-low sampling density. Using PSS and topography data as predictors of the soil properties, machine learning (ML) algorithms (support vector machine, random forest, and partial least squares) were trained for each sampling density and validated using an additional 20 independent soil samples. Differences between ML models or sampling densities were insignificant for a given soil property. However, the mean squared error (MSE) and the coefficient of determination (R^2) indicated that some models outperformed others. Models with an R^2 value above 0.5 were for P and pH with the 0.4 ha·sample^{-1} density and for P when the extra-low sampling design was applied. The definition of the evaluated ML algorithms does not consider the spatial location of the samples, which, from a mapping perspective, can create spatial inconsistencies; thus, to minimize this effect, an inverse distance smoothing window (SW) was applied to the predicted surfaces. The SW did not change predictions significantly, but often led to decreased R^2 and increased MSE values.

Keywords: *data fusion, machine learning, sensor calibration, soil fertility mapping*

The authors are solely responsible for the content of this paper, which is not a refereed publication. Citation of this work should state that it is from the Proceedings of the 16th International Conference on Precision Agriculture. EXAMPLE: Last Name, A. B. & Coauthor, C. D. (2024). Title of paper. In Proceedings of the 16th International Conference on Precision Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture.

Introduction

Within-field soil mapping is crucial to understanding, planning, and applying efficient, responsible, and accurate soil management practices. Soil sampling and its subsequent lab analysis is the most traditional approach for soil mapping and is often used for soil fertility assessment in agriculture (Gebbers, 2018). A single composite sample from a field might provide insights into its average fertility levels but not into its internal variability. Thus, grid and zone samplings are standard precision agriculture (PA) practices to evaluate the within-field variability. In the 2023 PA dealership survey, a long-term study conducted in the United States (Erickson and Lowenberg-DeBoer, 2023), PA retailers estimated that 51% of their local market area uses georeferenced soil sampling (i.e., grid or zone sampling).

Based on these numbers, precision agriculture practitioners use grid and zone sampling as a within-field fertility mapping approach. On the other hand, the U.S. dealerships involved in the above-mentioned survey estimated that 49% of the area did not use such a service. Such a high percentage of non-adoption could be attributed to the labor-intensive, time-consuming, and expensive nature of soil sampling, which induces farmers to opt for only a composite sample for the whole field or, in extreme cases, not to collect samples.

Therefore, there exists a challenge to develop and evaluate time- and cost-efficient methods for assessing the within-field soil variability to increase the adoption of PA strategies for best soil management practices (e.g., variable rate fertilizer application). Adamchuk et al. (2011) and Gebbers (2018) suggested the fusion of proximal soil sensors (PSS) as a potential solution to this challenge, leading others to evaluate this approach. Ji et al. (2019) assessed the potential of machine learning algorithms (ML) using different combinations of apparent electrical conductivity (EC_a), passive-γ-ray spectrometry, visible and near-infrared spectroscopy, and topography as predictors for soil chemical properties. These authors reported an improvement in the prediction when sensors were fused compared to when used individually. Saifuzzaman et al. (2021) fitted multivariate linear models to predict soil chemical properties using EC_a and topographic derivatives, obtaining similar findings as Ji et al. (2019) (i.e., there is a potential for data fusion to predict and map soil properties).

Despite the potential of PSS to improve agricultural soil maps, the signals captured by such sensors are complex and usually relate to a combination of processes occurring in the soil (Gebbers, 2018). Also, previous research results, such as by Ji et al. (2019) and Saifuzzaman et al. (2021), were obtained using sampling densities of approximately 0.25 ha·sample^{-1}, higher than commonly used by PA practitioners (1 ha·sample^{-1} - Erickson and Lowenberg-DeBoer, 2023). In addition, over the years, other geophysical techniques, such as ground penetrating radar (GPR), gained interest from the agricultural community, so their interactions with soil and its chemical properties must be investigated thoroughly.

Consequently, there is a need to explore the interactions at the source of the information provided by PSS and understand how the fusion of these data could provide better insights into soil spatial variability. The objectives of this study were to examine the relationship between proximal sensing data and soil properties, evaluate the feasibility of using the fusion of PSS and topography to improve the mapping of soil chemical properties and assess the effect of higher and lower sampling densities on the calibration model performance.

Material and Methods

Dataset description

A total of 128 samples [108 from a 0.4 ha·sample^{-1} sampling design (Fig. 1 – solid black circles) and an additional 20 independent validation samples (Fig. 1 – solid green diamonds)] were collected in 2022 from a 43-ha field in Central Alberta, Canada. Samples were removed from the original grid until a density of 3.5 ha·sample^{-1} was obtained (Fig. 1 – hollow blue squares), creating an extra-low density sampling design. All samples were collected under the same conditions and sent to the same laboratory for analysis. Samples collected from the topsoil layer (0-0.15 m) had their analysis results for plant-available potassium (K) and phosphorous (P), pH, and soil organic matter (OM) used to evaluate the prediction potential of PSS and topography. The above-described dataset is a subset of the one described by Karp et al. (2023a, 2024).

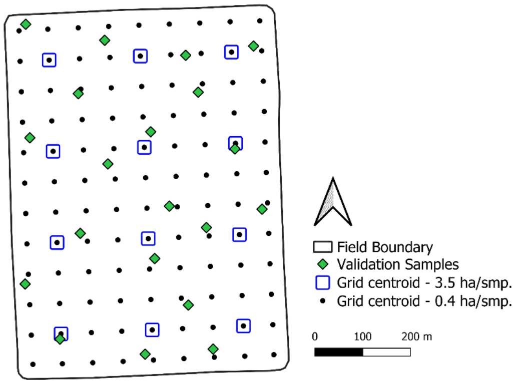


Fig. 1 Centroid locations for the original sampling density (0.4 ha·sample^{-1} – solid black circles), selected samples to create the extra low-density sampling design (3.5 ha·sample^{-1} – hallow blue squares), and validation samples (solid green diamonds) (Modified from Karp et al., 2024)

PSS data for soil resistance to penetration, EC$_a$, dielectric permittivity, and passive γ-ray for ^{137}Cs, ^{232}Th, ^{238}U, and ^{40}K were collected using the tools described in Table 1. The field's topography was assessed by creating its digital elevation model (DEM) using an unmanned aerial vehicle-mounted light detection and ranging (LiDAR) paired with a real-time kinematic correction (RTK) enabled global navigation satellite systems (GNSS) receiver. Hereafter, the five tools and variables described in Table 1 will be collectively referred to as "sources" and "predictor variables," respectively.

Due to timing constraints, such as weather conditions and the limited time window between harvest and first snowfall or last snowfall and crop development, the data from the different sources was not collected during the same season, but as follows: γ-ray in Spring 2019, ground penetrating radar (GPR) in Summer 2020, penetrometer in Summer 2021, electromagnetic induction (EMI) in Fall 2021, and DEM in Spring 2022. From 2019-2022, the field was cultivated with annual crops under a rotation of wheat, barley, and canola. Within this timeframe, no significant soil disturbing operations (e.g., field leveling, subsoiling) were performed, and the field was seeded and fertilized using uniform rates.

Processing for sensing data

EM38-MK2 and SoilOptix (Table 1) data did not require specific pre-processing strategies before further procedures or analysis, while the other three data sources did. All data manipulation and filtering are described in the following sub-sections.

Penetrometer Pre-Processing

The original data from the S600 penetrometer provides pressure (kPa) measurements from 0 to 0.6 m deep at 0.01 m intervals. Therefore, a total of 61 measurements are obtained for every sampled location. To reduce the number of variables and improve the data quality (i.e., reduce the effect of outliers), a mean boxcar with a window of 0.1 m (10 vertical measurements) was applied to the data. This process resulted in a dataset of 6 depth intervals of 0.1 m.

LiDAR Pre-Processing

LiDARMill (Phoenix LiDAR Systems, Austin, Texas, USA) was used to process the LiDAR data. This software automatically performed necessary data corrections, generated 788 points·m^{-2} georeferenced point cloud, and exported a 0.1-meter DEM raster. Finally, a custom R script (R Core Team, 2022) opened the DEM raster and calculated the fields' topography derivatives: slope, aspect, and curvature using the library *spatialEco* (Evans and Murphy, 2021).

Table 1 Description of the sensors, mapped variables, and collection settings adopted

Sensor Model	Manufacturer/ Provider	Technique	Swath width (m)	Density (points·ha^{-1})	Variables
S600 Penetrometer	Skok Agro (Vinnytsia, Vinnytsia Oblast-UA)	Cone Index	64	8	Resistance to penetration measured in pressure from 0 to 0.6 m (intervals of 0.01 m)
EM38-MK2	Geonic (Mississauga, Ontario-CA)	Electromagnetic Induction (EMI)	23	118	EC_a Shallow (0-0.75 m), EC_a Deep (0-1.5 m)
SoilOptix	SoilOptix (Tavistock, Ontario-CA)	Passive Gamma-Ray (γ-ray)	12	69	^{137}Cs Count, ^{232}Th Count, ^{238}U Count, ^{40}K Count, Count Rate (CR)
SIR-4000 (400MHz Antenna)	GSSI (Nashua, New Hampshire-USA)	Ground Penetrating Radar (GPR)	34	14,700	Soil Profile Amplitude through changes in dielectric permittivity
RECON-A	Phoenix LiDAR Systems (Austin, Texas-USA)	Light Detection and Ranging (LiDAR) + GNSS-RTK[a]	-	~7.8·10^{6}	Elevation[b]

[a]GNSS-RTK – Real Time Kinematic enabled Global Navigation Satellite System receiver; [b]Elevation – a product of processing the GNSS-RTK location with the measured LiDAR distances

Ground Penetrating Radar Pre-Processing

The SIR-400 GPR unit provided separate files for GNSS and GPR readings. To open and process the data, a custom Python script and the library *readgssi* (Nesbitt et al., 2022) were used. Below is a simplified description of the GPR processing; a detailed description can be found in Karp et al. (2023b).

Due to a higher collection frequency for the GPR than for the GNSS, linear interpolation was applied to the original coordinates, guaranteeing the georeferencing of all the sensor readings. In sequence, the GPR signal was processed by setting time-zero, using a "dewow" filter, removing background noise, applying a Hilbert transformation (calculates the signal envelope – instantaneous amplitude), and converting the signal travel time to relative depth (field estimated dielectric constant of 12.79).

The GPR unit provided 512 vertical readings from the soil at every sampling location. Thus, the GPR instantaneous amplitude was subjected to a boxcar median with a 0.1 m window size. The maximum processing depth was set to 2 m, resulting in 20 depth layers. Since the density of the GPR data was very high within the collection transect (1 sample every 0.02 m – hence the high collection density in Table 1), a boxcar median was also applied in the direction of travel for the data collection. A 5 m distance between consecutive sampling points was achieved using a 250 samples window.

General PSS data filtering procedure

After completing the specific processing for the individual data sources, all PSS data were filtered to reduce the effect of outliers and maximize the data quality. The filtering procedure followed the steps suggested by Karp et al. (2022): (1) project the dataset to a custom localized Cartesian coordinate system, (2) identify and apply a position offset (e.g., the distance between the GNSS receiver and sensor), (3) operational filtering (i.e., removal of maneuvers, abrupt speed changes), (4) global and local statistical filtering.

Dataset co-location

To investigate the predicting capabilities of soil chemical properties using PSS and topography data, all the data must be co-located. Two approaches were adopted for the data co-location: one focused on building a dataset for training the predictive algorithms, and the other on predicting the spatial distribution of the soil property.

During the soil sampling activity, a GNSS receiver was used to record the location where the core samples were collected. At each core location, three subsamples were taken within a 5-meter radius. This intrinsic characteristic of the sampling method defined the first co-location method. A 5-meter buffer was applied to the recorded core locations, and the median of PSS or topography observations within the buffered area was calculated and attributed to the corresponding sampling location. The final dataset comprised 41 columns: 6 depth intervals of soil resistance to penetration from the penetrometer, 20 depth intervals of instantaneous amplitude from GPR, two depth ranges of EC_a from EMI, five count rates (total count rate plus the four separate nuclides) from γ-ray, four from topography (elevation, curvature, aspect, and slope), and the soil analysis for P, K, pH, and OM.

The above-described approach potentially minimizes issues with the change of support and co-location inaccuracies; however, it does not provide a continuous surface. Thus, in the second approach, inverse distance weighting (IDW) was used to interpolate the PSS data to a 15-meter resolution raster. For the topography data, the 0.1-m raster was downscaled to the same 15-meter raster using the median resampling method from *gdalwarp* (GDAL/OGR contributors, 2024). This approach resulted in a 37-band raster containing only the predictor variables.

Data preliminary analysis and predictive modeling

The descriptive statistics for PSS, the two soil sampling densities, and validation samples were calculated in a preliminary data analysis. The original 0.4 ha·sample^{-1} co-located dataset was used to study the relationships between predictors and the soil properties by correlation analysis.

Thereafter, all the data were standardized to a zero mean and a unit variance for homogeneity purposes. Partial least squares (PLSR; Wold et al., 1983), support vector machine (SVM; Platt, 2000), and random forest (RF; Breiman, 2001) algorithms were evaluated using training data from the co-located 0.4 and 3.5 ha·sample^{-1} sampling designs. The three algorithms were implemented with a customized Python script using the library *scikit-learn* (Pedregosa et al., 2012). The Python library *Optuna* was used to tune the model parameters individually for a given soil variable, ML, and sampling density.

Most ML models benefit from larger datasets, and when exposed to a small number of observations and many predictors, they can overfit the training dataset (i.e., reduce the capability of generalizing the model). While a sampling density of 0.4 ha·sample^{-1} was available for this study site, coarser sampling designs are more common among PA practitioners. According to Erickson and Lowenberg-DeBoer (2023), two common barriers to the adoption of PA are related to the farm income and the PA service costs. Even though the 3.5 ha·sample^{-1} sampling density provided only 12 samples, which might affect model performance, it broadened the discussion and produced realistic and practical results. From an economic perspective, using the original sampling density could defeat the option of collecting PSS and elevation data.

None of the three ML algorithms mentioned above includes a spatial structure for the data, which is likely to be an issue when using the trained models to predict a continuous surface. Therefore, the application of an inverse distance weighted (IDW) moving window was evaluated.

A window size and a matrix of weights are required to define the moving window. The window size limits the neighborhood of cells used to estimate the value at the center of the window. An estimate of the range of spatial autocorrelation was used as a basis to define the window size, assuming a circular shape and isotropy. Low-density sampling designs often will not provide enough information to obtain variogram estimates with traditional methods, so the approach proposed by Karp et al. (2024) was adopted. When a flat, pure-nugget effect variogram model was calculated, the minimum distance between sampling locations determined the window size.

The formula 1/(distance to window center)^2 determined the weights. Map cells with a distance from the center of the window, greater than half the range of spatial autocorrelation, were removed, yielding a circular neighborhood. Finally, the weights were normalized so that the sum of their values was each time equal to 1.

Predictive modeling comparison

Mean Squared Error (MSE) and the coefficient of determination (R^2) were used to assess the predictive results' performance. For each soil property, three ML algorithms, two sampling densities, and two predicted surface treatments ("No Spatial Smoothing" and "IDW Moving Window") were evaluated, resulting in 12 different predictive results per soil property. These results were assessed using 20 independent validation samples for a given soil property. The squared errors from the results were compared through a pairwise Levene's test; when the null hypothesis was rejected (homogeneity of variances), meaning there was heterogeneity in the variance of the squared errors, the two evaluated prediction results were considered different.

The effect of ML algorithms on prediction accuracy was evaluated independently for each sampling design and a given soil property. A multi-objective decision-making logic was adopted to guarantee the selection of the most robust ML algorithm:

1. Models that presented statistically significantly lower errors were selected. If no statistically significant difference was observed, all models were selected.
2. If only one model was selected in Step 1, go to Step 5. Otherwise, MSE was standardized to a scale between 0 and 1.
3. A score was calculated using the formula R^2 + (1- standardized MSE) and ranked in a descending order.
4. The model with the highest score was selected as the best predictor. The selected model was considered the most robust ML algorithm for the given dataset.

Using R^2 and MSE simultaneously avoids selecting models with low MSE and low R^2 and high R^2 and high MSE while selecting accurate models that best explain the variance of the soil property in the validation samples.

The most robust models for the two sampling densities were then compared, analyzing the effect of sampling density on the prediction of soil chemical properties. The IDW Moving Window effect was evaluated in sequence for a given density. Finally, thematic maps were generated for predicted surfaces and compared to surfaces obtained through ordinary kriging interpolation of the 0.4 ha·sample^{-1} sampling design. All data, interpolation, and statistical analysis were performed using custom scripts written in the R language, and all standardized data and predictions were back transformed to report the results.

Results and Discussion

Descriptive Statistics

The descriptive statistics for the validation samples and the original and extra-low soil sampling designs are presented in Table 2. Similar standard deviations (SD), means, and medians are observed for the original (0.4 ha·sample^{-1}) and extra-low (3.5 ha·sample^{-1}) designs for a given soil property, indicating that there is a good overall representation of the underlying surface even with only 12 samples. The inline histograms from these two sampling densities differ, though, an expected response due to the reduction of almost 90% of samples. Note that pH and OM presented a smaller variance than P and K for both sampling designs.

The means and medians for the validation samples are lower for K, P, and OM than for both sampling designs, whereas they are slightly higher for soil pH. The SDs of the validation samples are lower than that of the grid sampling for K and slightly higher for P, pH, and OM. Except for pH, the inline histograms for the validation samples are similar to the 0.4 ha·sample^{-1} design. Despite these slight differences, the validation samples capture a similar variability as the grid samples, an important characteristic to guarantee the validity of further analysis using this dataset.

Table 2 Descriptive statistics for soil testing results for plant-available potassium (K) and phosphorous (P), pH, and soil organic matter (OM) from two sampling densities and validation samples. (Modified from Karp et al., 2024)

Density	Grid Samples						Validation Samples				
	n[a]	Variable	Mean	Standard Deviation	Median	Histogram	n[a]	Mean	Standard Deviation	Median	Histogram
0.4 ha·sample^{-1}	108	K (ppm)	133.9	37.6	130.5		20	128.0	31.5	122.5	
		P (ppm)	34.0	14.8	33.5			29.3	18.4	24.0	
		pH (-)	7.37	0.49	7.30			7.58	0.79	7.75	
		OM (%)	7.24	0.69	7.20			6.72	0.70	6.75	
3.5 ha·sample^{-1}	12	K (ppm)	127.8	40.7	136.5		20	128.0	31.5	122.5	
		P (ppm)	31.3	15.4	33.0			29.3	18.4	24.0	
		pH (-)	7.39	0.51	7.20			7.58	0.79	7.75	
		OM (%)	7.3	0.64	7.55			6.72	0.70	6.75	

[a] n: number of samples

Similarly, Table 3 presents the descriptive statistics for one of the variables for each PSS before and after the general filtering procedure. The filtering procedure consistently reduced the SDs and differences between the means and medians of each data source. The comparison of the inline histograms only indicated minor changes in the data distribution, suggesting a successful removal of outliers while maintaining the integrity of the data distribution.

Table 3 Example of descriptive statistics for the elevation raster data and one variable from each PSS before (raw) and after applying the filtering procedure steps from Karp et al. (2022) (a hyphen indicates that the filtering procedure was not applied to that dataset)

Variable	Raw data				Filtered data			
	Mean	Standard Deviation	Median	Histogram	Mean	Standard Deviation	Median	Histogram
Resistance to penetration (kPa) from 0.01-0.10 m	219.2	215.9	172.0		174.6	63.6	168.3	
Electromagnetic Induction EC$_a$ (mS·m^{-1}) from 0-0.75 m	248.4	4.7	247.2		248.0	3.3	247.2	
Ground Penetrating Radar Instantaneous amplitude (-) 0.00-0.10 m	250531.7	148691.5	214862.9		223260.4	60323.4	209760.1	
γ-ray ^{40}K (count rate)	294.2	148.5	283.5		286.5	35.6	285.2	
Elevation (m)	1022.6	4.2	1022.6		-	-	-	-

A lower SD and mean ratio indicate a lower elevation and EC$_a$ variance than other variables. The evaluation of maps from these two data sources (not included in this paper) still showed that these variables represented this field's known spatial variability. Such observation highlights the importance of standardizing the dataset to zero mean and unit variance, as ML algorithms can be sensitive to the magnitude and variance of the data, which could result in reduced importance of t the two variables mentioned above.

Representative training and validation datasets are essential to guarantee the model's validity and analysis of the results. The descriptive statistics for the same variables from the different data sources are presented for the surface resulting from the IDW interpolation of the data sources and for the 3.5 ha·sample^{-1} training dataset. Despite some slight differences, the descriptive statistics for the training dataset and interpolated surfaces (Table 4) are similar to the ones from the "Filtered data" from Table 3 ("raw data" for elevation). The similarity among the descriptive statistics from these datasets indicates that no substantial changes were induced through the co-location and interpolation procedures, and reliable datasets were generated to achieve the objectives of this study.

Table 4 Example of descriptive statistics from the 3.5 ha·sample^{-1} co-located training dataset and interpolated surface (15-meter raster) for elevation and one variable from each PSS

Variable	Training Dataset (3.5 ha·sample^{-1})				Interpolated Surface			
	Mean	Standard Deviation	Median	Histogram	Mean	Standard Deviation	Median	Histogram
Resistance to penetration (kPa) from 0.01-0.10 m	171.4	56.6	167.5		186.8	59.5	187.3	
Electromagnetic Induction EC_a (mS·m^{-1}) from 0-0.75 m	248.0	3.4	247.0		247.4	2.8	246.3	
Ground Penetrating Radar Instantaneous amplitude (-) 0.00-0.10 m	202015.3	63791.7	190354.6		238203.8	45946.2	226345.0	
γ-ray ^{40}K (count rate)	278.5	33.4	283.1		289.0	22.8	290.5	
Elevation	1023.1	3.5	1022.5		1022.4	4.5	1023.2	

Correlation analysis

The collinearity in the predictor variables and the relationships between predictors and the soil chemical properties were assessed on the dataset with 0.4 ha·sample^{-1} density. A visual inspection of the histograms in Table 3 allows to see that even after filtering, the distribution for some of the predictors did not look like a normal distribution. Thus, Spearman's correlation coefficient was used instead of Pearson's. The resulting correlations are presented in Fig. 2Fig. 3.

The correlations among predictors (Fig. 2) could lead to a lengthy discussion from a geophysical and engineering-focused perspective (Karp et al., 2023b). Thus, in the present study, the content of Fig. 2 is discussed and interpreted from a modeling perspective.

Signal responses from different soil layers belonging to the same PSS are often correlated, for which subsequent intervals provide higher correlation values (e.g., Penetrometer 0.11 – 0.20 and 0.21 – 0.30 m). Such behavior is ‘as expected’ since spatial correlation is not limited to its most explored dimension (2D). Across data sources, all predictors significantly correlate to one or more predictor variables.

The observed correlation within and across data sources can be seen as a multicollinearity issue, which might result in unstable estimation of coefficients and variance inflation, consequently affecting the models’ predicting capabilities (Allen, 1997). This known limitation in data fusion is commonly handled using feature selection approaches (Ji et al., 2019; Lachgar et al., 2024) to remove variables that do not contribute to or negatively impact the models’ performance. The present study avoided the removal of predictor variables and entirely relied on the potential of some of the chosen ML algorithms to overcome this limitation. For example, PLSR reduces the dataset dimension and correlations among predictors through latent variables that maximize the explanation of the target variable. RF uses random sampling of variables and observations to reduce the overfitting of the model, which can reduce the effect of multicollinearity. Without a process that can minimize collinearity among predictors, SVM is the most vulnerable to this effect.

The correlations between the predictors and soil properties are presented separately in Fig. 3 to facilitate the visualization and analysis. Except for ^{238}U and slope, all other variables demonstrated a significant correlation with at least one soil property. The 0.21-0.3 m resistance to penetration and γ-ray count rates for ⁴⁰K and total count rate (CR) presented significant correlation with all soil properties. pH and P correlate to the greatest number of predictors, 28 and 10, respectively. These two soil properties also present absolute correlation values above 0.5. The highest absolute correlation for pH, P, K, and OM are with "Penetrometer 0.31-0.40" (-0.59), "Penetrometer 0.21-0.30" (0.55), "Elevation" (0.44), and "γ-ray ^{40}K" (-0.32), respectively.

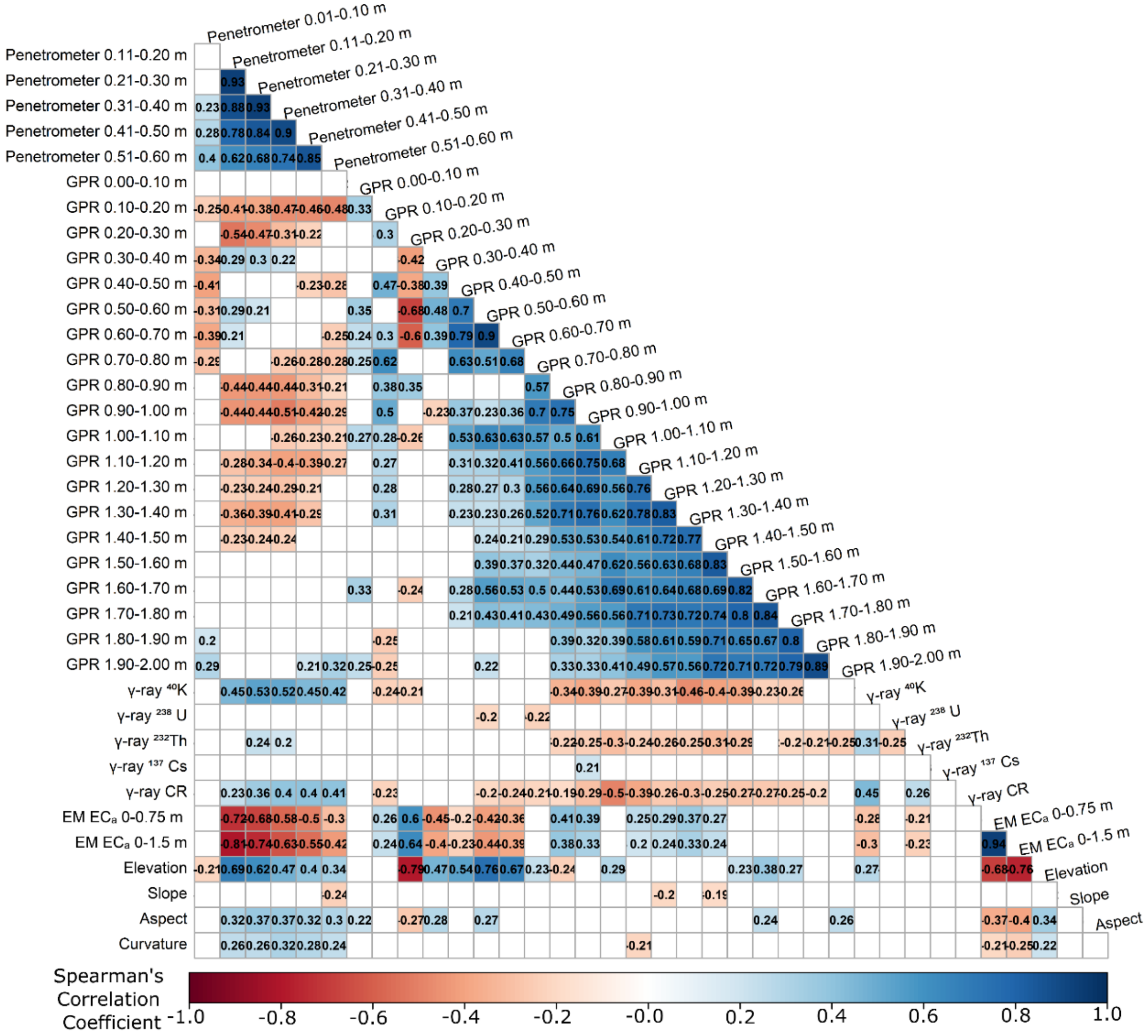


Fig. 2 Spearman's correlogram for all sensors and topography data. Empty cells represent the non-significant correlation at a significance level of 0.05. The darker the cell color, the stronger the correlation (positive correlation – blue; negative correlation – red). GPR – ground penetrating radar; ECₐ- apparent electrical conductivity; EM – Electromagnetic Induction sensor; CR – count rate.

No complete agreement between a predictor variable and soil properties was identified. Also, each soil property relates to the data sources differently. For instance, pH significantly correlates to multiple variables from GPR, while OM to none. In contrast, all soil properties correlated to at least three variables from the penetrometer. However, the correlation is stronger (absolute values above 0.5 for some depth intervals) for pH and P and weaker for OM and K (highest absolute value is 0.33). These observations lead to the conclusion that no single sensor can measure or predict all soil properties, a well-known behavior when mapping soil variability with PSS (Adamchuk et al., 2011; Gebbers, 2018).

Upon further inspection of the interpolated surfaces, all predictors clearly defined different aspects of the known variability of the study field, except for ^{238}U and ^{137}Cs, whose maps mainly presented a poorly structured spatial distribution. Since ^{238}U did not show a significant correlation with any soil property, and while analyzing the γ-ray dataset from this field, Karp et al. (2023b) observed pure nugget effect experimental variograms for ^{137}Cs and ^{238}U; these two nuclides were not used for the model training and predictions.

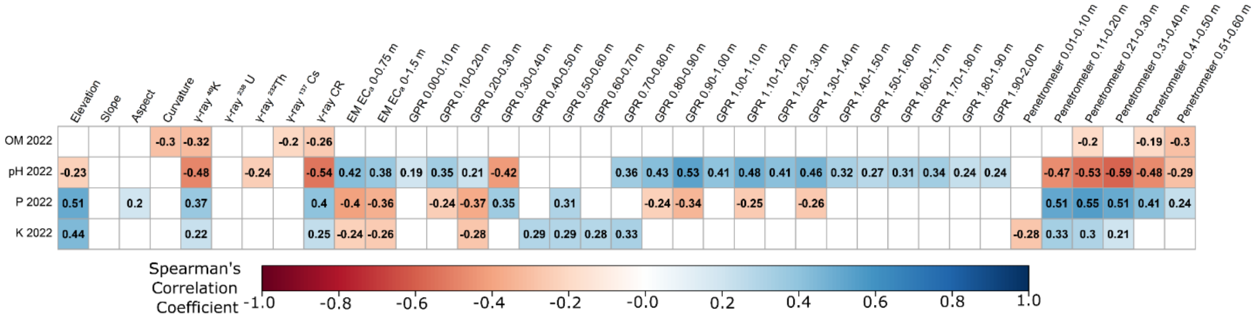


Fig. 3 Spearman's correlogram matrix for soil properties, all sensors, and topography data. Empty cells represent the non-significant correlation at a significance level of 0.05. The darker the cell color, the stronger the correlation (positive correlation – blue; negative correlation – red). GPR – ground penetrating radar; EC$_a$- apparent electrical conductivity; EM38 – Electromagnetic Induction sensor; CR – count rate.

Calibration Results

The model parameters for a given ML algorithm, sampling density, and soil chemical property were tuned, and the best parameters were used to train the calibration models successfully.

Effect and performance of ML algorithms for a given sampling density and soil property

The MSE from the ML algorithms were compared using Levene's test. No statistical significance (α = 0.05) was identified for a given soil property and sampling design, meaning homogeneity of variances in the squared errors from the three ML algorithms. A comparison of MSE and R^2 from Fig. 4 suggests that the multi-objective decision-making approach described above demonstrated to be an effective approach to select the most robust ML algorithms (red-bordered bars in Fig. 4). For example, when using the 3.5 ha·sample^{-1} training dataset for predicting K (Fig. 4a), RF and SVM resulted in very similar MSE, while RF was selected due to its higher R^2.

SVM and RF emerged as the most robust algorithms for 3 out of 4 soil properties for the original and extra-low density sampling designs, respectively (Fig. 4a, c-d). For P, PLSR outperformed the other models for both sampling densities (Fig. 4b). Ji et al. (2019) compared the performance of ML algorithms (including PLSR, RF, and SVM) to predict soil properties using the fusion of soil γ-ray, reflectance from visible and near-infrared spectra, EC$_a$ from EMI, and elevation. The results presented by Ji et al. (2019) indicate that PLSR was often outperformed by SVM and RF, which aligns with the results observed in Fig. 4.

The SVM never emerged as the most robust algorithm for the extra-low density sampling design; as previously mentioned, the higher susceptibility of this algorithm to collinearity might be contributing to this result. Thus, future research should evaluate how feature selection could change the observed results.

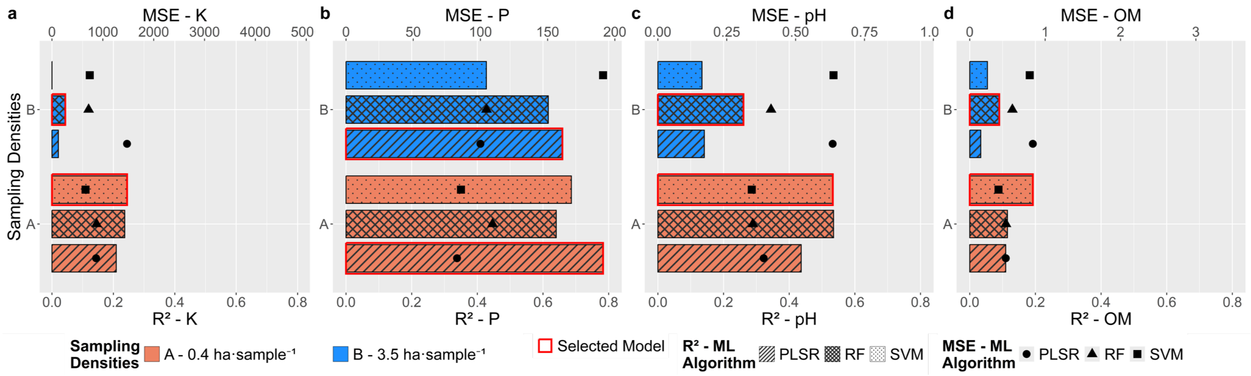


Fig. 4 Bar-dot plot for the coefficient of determination (R^2; bars) and mean squared error (MSE; points) for the 20 validation samples comparing partial least squares (PLSR; hatched bars and circle-shaped points), random forest (RF; crosshatched bars and triangle-shaped points), and support vector machines (SVM; bars with small dots and square-shaped points) as prediction algorithms for plant available potassium (K), phosphorus (P), pH, and soil organic matter (OM) for a given sampling density. A red border around a bar indicates the selected model for that sampling density.

Effect and performance of sampling density on the prediction of the soil properties

Figs. 5a-d present a focused analysis of the effect of sampling density on predicting soil chemical properties. The R^2 is reported in the boxes on the left side of the panel, while MSE is on the right. Even though Levene's test results in Figs 5a-d (uppercase letters following the MSE values) did not reveal significant differences in the variance of squared residuals between the two sampling densities, the performance metrics indicated that calibration models using the original sampling density consistently outperformed the extra-low-density.

According to the results presented in Figs 5a-d, only the predictions for P yielded R^2 above 0.5 for both sampling densities (Fig. 5b). For pH (Fig. 5c), the R^2 for the 0.4 ha·sample^{-1} model was 0.53, while 0.26 for the lower density design. These results for pH already suggest a gain in the percentage of the variability that the sensor's fusion can explain when including more samples. This result is 'as expected' since adding more samples improves parameter tuning and model predictions. Also, with the addition of closer samples (higher sampling density), spatial autocorrelation among sampling locations becomes more representative in the training dataset, which is not accounted for in the evaluated ML algorithms but affects the model predictions – the model overfits to the dataset (Hengl et al., 2018). The models for the original and lower density sampling designs accounted for less than 25% and 10%, respectively, of the variability in the validation samples for K (Fig. 5a) and OM (Fig. 5d).

The lower R^2 observed for OM (Fig. 5d) could be attributed to the lower variance of this soil property in this field (Table 2). While pH also presented a lower variance (Table 2), a stronger relationship between the penetrometer variables and pH was observed (Fig. 3), which might have contributed to the higher R^2 for this soil property. In contrast to OM and pH, K presented a higher variance (Table 2). Thus, the resulting R^2 for K suggests that calibration models trained with the fusion of the five different data sources could not explain more than 24% and 4% of the variability of this soil property in the validation samples when using the 0.4 and 3.5 ha·sample^{-1} sampling densities, respectively.

The above observations regarding the modes' performance can also be observed in the predicted surfaces. Fig 6 compares thematic maps from the data fusion calibration models with the ordinary kriging (OK) interpolation of the 0.4 ha·sample^{-1} sampling density. A visual comparison of the surfaces for P and pH revealed that maps originated from ML models for both sampling densities (P – Figs. 6g and i; pH – Figs. 6l and n) agree strongly with their respective OK maps (P – Fig. 6f; pH – Fig 6k). In contrast, such an agreement is weaker for K and OM, with the original sampling design models predicting surfaces closer to OK rather than for the extra-low sampling design models. Overall, the observed agreements between the ML predicted surfaces with OK interpolation suggests that the co-location procedures for the training and prediction dataset were effective.

Considering that the fused dataset contained 37 variables, some might negatively affect the model's predictability, as could the different collection dates for the data sources. Again, this indicates that future work should explore feature selection approaches to reduce the number or better select predictors or data sources.

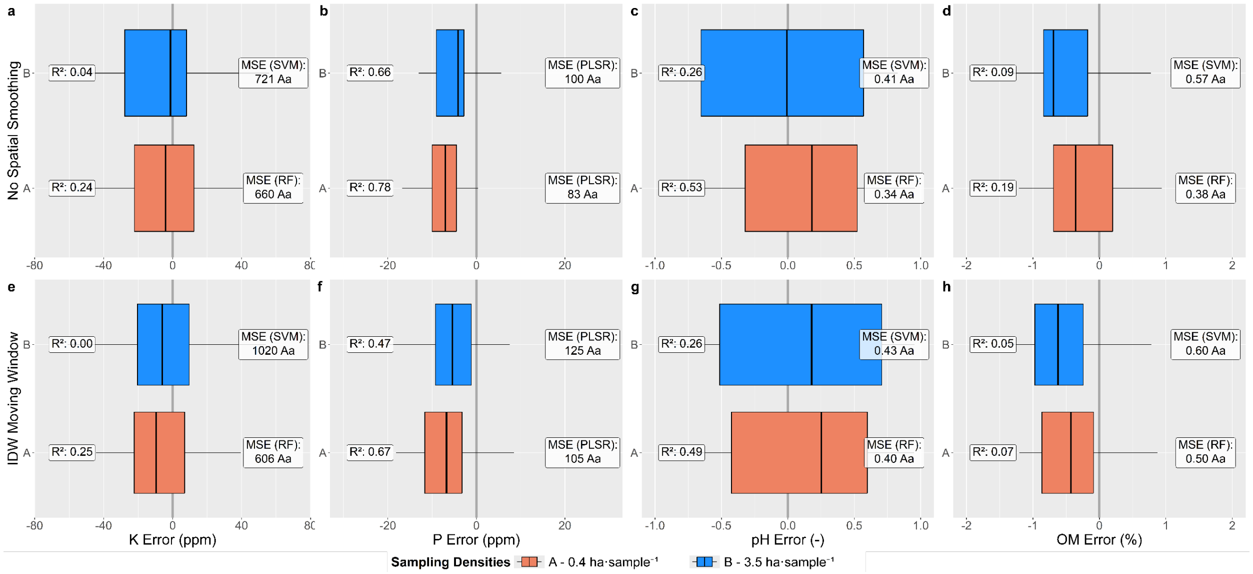


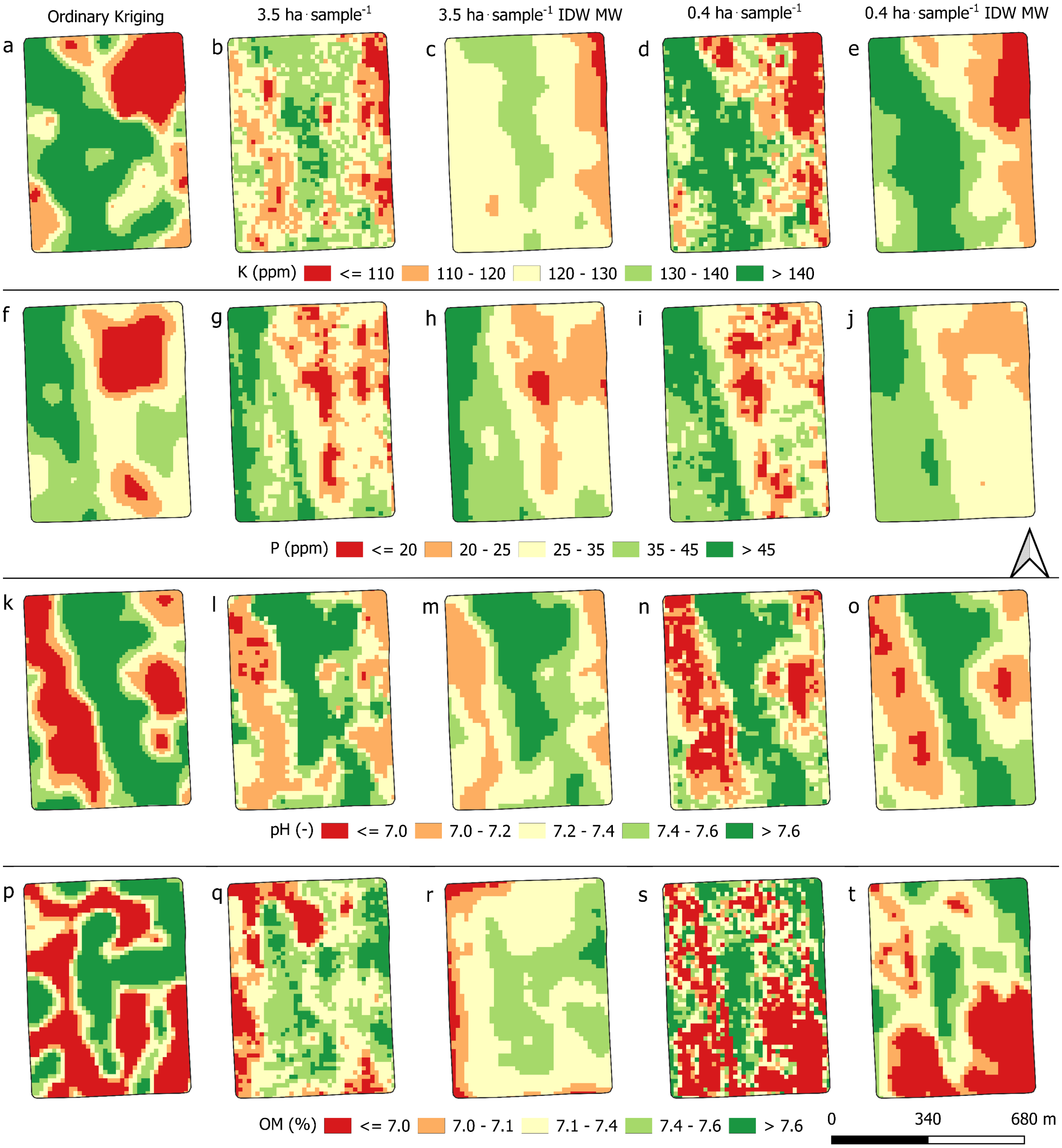
Fig. 5 Box plots for the prediction errors for plant-available potassium (K), phosphorus (P), pH, and soil organic matter (OM). Results from the predictions based on the fusion of all data sources for two sampling densities, 0.4 and 3.5 ha·sample^{-1}, are indicated by a capital letter and a color. While partial least squares (PLSR), random forest (RF), and support vector machines (SVM) were tested for each scenario, only the best-performing algorithms are presented (refer to Fig. 4 for all algorithms). MSE followed by different uppercase letters differ significantly at α = 0.05 within the panel (between sampling densities), and different lowercase letters for the same sampling density but between "No Spatial Smoothing" and "IDW Moving Window"

Effect of an IDW smoothing approach in the surface predicted by a non-spatial ML algorithm

Although research results have supported that adding coordinates, sampling distances, and neighboring observations as covariates can improve the prediction capability of the ML algorithms (Hengl et al., 2018; Pereira et al., 2022; Sekulić et al., 2020; Talebi et al., 2022), such approaches were not explored in the current study, as a focus was given to calibrating the fused dataset.

None of the evaluated ML algorithms accounted for the spatial component in the data; the predicted surfaces can present spatial outliers. This behavior can be observed in the maps for "3.5 ha·sample^{-1}" and "0.4 ha·sample^{-1}" in Figs 6. From a practical perspective, such spatial inconsistency in the maps might affect prescription maps. Therefore, an IDW smoothing approach was evaluated, and the results are presented in Fig. 5e-h. Since none of the MSE values were followed by a different lowercase letter, Levene's test did not indicate a significant difference (α = 0.05) in the MSE after applying the IDW Smoothing Window for a given sampling density. However, the smoothing approach often worsens the performance metrics (Figs. 5e-h) compared to the standalone model predictions (Figs. 5a-d). A visual comparison of the smoothed maps and standalone predictions (Fig. 6) suggests that the proposed smoothing approach reduced spatial inconsistencies.

For a reduction of 90% in the number of samples, using ML algorithms and the fusion of PSS and topography data presented some potential to predict the spatial variability of P and pH when using extra-low sampling density. From the perspective of a PA practitioner, such results might be more appealing than those obtained for the 0.4 ha·sample^{-1} design. It is important to note that these results are applicable for the specific experimental site. The effect of sampling density on the model performance will vary for different fields and soil properties, as it depends on the spatial structure of the specific site and dataset. Thus, the results presented above should be generalized with care. Also, there are still limitations on the prediction of OM and K, which are also crucial for fertility management purposes. Therefore, further evaluations should be performed using different combinations of data sources and adding additional data sources that can help predict these two chemical properties.



Ordinary Kriging
3.5 ha·sample^{-1}
3.5 ha·sample^{-1} IDW MW
0.4 ha·sample^{-1}
0.4 ha·sample^{-1} IDW MW
a
b
c
d
e
K (ppm) 🟥 <= 110 🟧 110 - 120 🟨 120 - 130 🟩 130 - 140 🟩 > 140
f
g
h
i
j
P (ppm) 🟥 <= 20 🟧 20 - 25 🟨 25 - 35 🟩 35 - 45 🟩 > 45
k
l
m
n
o
pH (-) 🟥 <= 7.0 🟧 7.0 - 7.2 🟨 7.2 - 7.4 🟩 7.4 - 7.6 🟩 > 7.6
p
q
r
s
t
OM (%) 🟥 <= 7.0 🟧 7.0 - 7.1 🟨 7.1 - 7.4 🟩 7.4 - 7.6 🟩 > 7.6
0
340
680 m

Fig. 6 Thematic maps for plant-available potassium (K; a-e), phosphorus (P; f-j), pH (k-o), and soil organic matter (OM; p-t) from the interpolation of the 0.4 ha·sample^{-1}using ordinary kriging (first column of maps), and calibration model predictions for the 3.5 ha·sample^{-1} (second and third columns) and 0.4 ha·sample^{-1} (fourth and fifth columns) before and after the "IDW Moving Window" (IDW MW) was applied

Conclusion

Although significant correlations between soil chemical properties and PSS and topography data were observed, no complete agreement was observed. The complex relationship between the properties measured by the sensors and the soil variables, as previously reported by other researchers, is a possible explanation for this behavior. These results indicate a potential benefit of fusing the data sources.

The machine learning algorithms evaluated did not present statistically significant differences when the squared residuals were compared using Levene's test. However, the use of a multi-objective decision making logic highlighted some differences between the models and presented to be an effective approach to select the best predictor for the two training datasets (0.4 and 3.5 ha·sample^{-1}).

No statistical difference was observed when comparing the residuals of the most robust ML algorithms from the two sampling densities evaluated. However, the performance metrics indicated that the higher-density dataset provided better predictions. The models for P and pH provided better results than those for OM and K, which models (especially when using the lower-density design) did not account for more than 25% of the variability in the validation dataset. Overall, a visual agreement between some of the predicted surfaces and OK interpolation (more evident for P and pH) was observed suggesting that the co-location procedures adopted for training and prediction were effective.

The ML algorithms evaluated did not consider the spatial component in the data, creating spatial outliers in the predicted surfaces. To overcome this limitation, an IDW-based moving window was evaluated, but while it reduced spatial inconsistencies, it slightly worsened the model performance metrics.

The results presented the potential of calibrating PSS and topography data fusion to predict soil chemical properties. However, this needs to be further explored, especially regarding the different combinations of the data sources.

Acknowledgments

We thank the Canadian Agri-Food Automation and Intelligence Network (CAAIN), Mitacs, and Telus Agriculture for funding the project and data collection. We also thank SoilOptix, Inc. (Tavistock, ON, Canada) for providing the complete gamma-ray data. This research is part of the project 'Agricultural Multi-Layer Data Fusion to Support Cloud-Based Agricultural Advisory Services' funded by Mitacs through the Mitacs Accelerate program.

References

Adamchuk, V. I., Viscarra Rossel, R. A., Sudduth, K. A., & Lammers, P. S. (2011). Sensor Fusion for Precision Agriculture. In *Sensor Fusion - Foundation and Applications*. InTech. https://doi.org/10.5772/19983

Allen, M. P. (1997). The problem of multicollinearity. In *Understanding Regression Analysis* (pp. 176–180). Boston, MA: Springer US. https://doi.org/10.1007/978-0-585-25657-3_37

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Erickson, B., & Lowenberg-DeBoer, J. (2023). 2023 Precision Agriculture Dealership Survey. Department of Agronomy and Agricultural Economics, Purdue University. https://ag.purdue.edu/digitalag/_media/croplife-purdue-precision-dealer-report-2023.pdf

Evans, J. S., & Murphy, M. A. (2021). spatialEco. https://github.com/jeffreyevans/spatialEco

GDAL/OGR contributors. (2024). GDAL/OGR Geospatial Data Abstraction software Library. https://doi.org/10.5281/zenodo.5884351

Gebbers, R. (2018). Proximal soil surveying and monitoring techniques (pp. 29–78). https://doi.org/10.19103/AS.2017.0032.01

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *2018*(8). https://doi.org/10.7717/peerj.5518

Ji, W., Adamchuk, V. I., Chen, S., Mat Su, A. S., Ismail, A., Gan, Q., et al. (2019). Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study. *Geoderma*, *341*(July 2017), 111–128. https://doi.org/10.1016/j.geoderma.2019.01.006

Karp, F. H. S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2023a). Comparative study of interpolation methods for low-density sampling. In *Precision agriculture ’23* (Vol. 34, pp. 563–569). The Netherlands: Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-947-3_71

Karp, F. H. S., Adamchuk, V., Dutilleul, P., & Melnitchouck, A. (2024). Comparative study of interpolation methods for low-density sampling. *Precision Agriculture*. https://doi.org/10.1007/s11119-024-10141-0

Karp, F. H. S., Adamchuk, V. I., Melnitchouck, A., Allred, B., Dutilleul, P., & Martinez, L. R. (2023b). Validation And Potential Improvement of Soil Survey Maps Using Proximal Soil Sensing. *Journal of Environmental and Engineering Geophysics*, *28*(1), 45–61. https://doi.org/10.32389/JEEG22-018

Karp, F. H. S., Adamchuk, V., Melnitchouck, A., & Dutilleul, P. (2022). Optimization of Batch Processing of High-Density Anisotropic Distributed Proximal Soil Sensing Data for Precision Agriculture Purposes. In *Proceedings of the 15th International Conference on Precision Agriculture* (p. unpaginated, online). Monticello, IL: International Society of Precision Agriculture. https://www.ispag.org/proceedings/?action=abstract&id=8792&title=Optimization+of+Batch+Processing+of+High-density+Anisotropic+Distributed+Proximal+Soil+Sensing+Data+for+Precision+Agriculture+Purposes

Lachgar, A., Mulla, D. J., & Adamchuk, V. (2024). Implementation of Proximal and Remote Soil Sensing, Data Fusion and Machine Learning to Improve Phosphorus Spatial Prediction for Farms in Ontario, Canada. *Agronomy*, *14*(4). https://doi.org/10.3390/agronomy14040693

Nesbitt, I., Simon, F.-X., Hoffmann, F., Paulin, T., & teshaw. (2022). readgssi: an open-source tool to read and plot GSSI ground-penetrating radar data. https://doi.org/10.5281/ZENODO.5932420

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://arxiv.org/abs/1201.0490

Pereira, G. W., Valente, D. S. M., de Queiroz, D. M., Santos, N. T., & Fernandes-Filho, E. I. (2022). Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. *Precision Agriculture*, *23*(4), 1189–1204. https://doi.org/10.1007/s11119-022-09880-9

Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in Large-Margin Classifiers*. Cambridge: MIT Press.

R Core Team. (2022). R: A Language and Environment for Statistical Computing. Vienna, Austria. https://www.r-project.org/

Saifuzzaman, M., Adamchuk, V., Biswas, A., & Rabe, N. (2021). High-density proximal soil sensing data and topographic derivatives to characterise field variability. *Biosystems Engineering*, *211*, 19–34. https://doi.org/10.1016/j.biosystemseng.2021.08.018

Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, *12*(10), 1–29. https://doi.org/10.3390/rs12101687

Talebi, H., Peeters, L. J. M., Otto, A., & Tolosana-Delgado, R. (2022). A Truly Spatial Random Forests Algorithm for Geoscience Data Analysis and Modelling. *Mathematical Geosciences*, *54*(1), 1–22. https://doi.org/10.1007/s11004-021-09946-w

Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method (pp. 286–293). https://doi.org/10.1007/BFb0062108