ORIGINAL ARTICLE

Samut S

The comparison of Linear Mixed Model and Fuzzy clustering to predict soil CEC and pH using proximal and remote geophysical data over infertile soils of Northeast Thailand

Khongnawang T.[1*] | **Srihabun P.**[2*]

[1]Land Development Regional Office 10

[2]Land Development Regional Office 5

Correspondence
Email:

Present address
[†]Samut Songkram Land Development Station, Land Development Regional Office 10, Land Development Department, Samut Songkram Province, 75000, Thailand

Funding information
Supported by Soil Salinity Improvement Project, Land Development Regional Office 5

The spatial prediction of soil properties is necessarily for agricultural purposes. Especially for the information of the cation exchange capacity (CEC - cmol(+) kg^{-1}) and the soil pH which are the key to soil improvement. However, the traditional method in mapping soil properties is known as a constraint with soil chemical properties. To avoid this, the digital soil mapping (DSM) technique using two potential models was taken into account. The first is the direct comparative linear mixed model (LMM) and the second is the k-mean clustering algorithm. We compare the performance of these two models by determining on the mean square prediction error (MSPE, σ^2) when using various digital data, including proximal data from $\gamma - ray$ spectrometer (i.e. K, U, Th and TC), EM38 (i.e. EC_{ah} and EC_{av}) and remote sensed data from Leaf Area Index (LAI) and Digital Elevation Model (DEM) in predicting CEC and pH at various depths, including topsoil (0-0.3 m), subsurface (0.3-0.6 m) and subsoil (0.6-0.9 m). The LMM prediction results for the topsoil and subsurface CEC were best with proximal (σ_p^2=

1.93 and 2.60, respectively), while the subsoil was the combination of proximal and remote sensed data (σ_p^2 = 6.05). In terms of soil pH, the regression model was significant with the proximal plus remote (σ_p^2 = 0.18), while the subsurface and the subsoil were best with either proximal (σ_p^2= 0.18 and 0.31, respectively). On the other hand, the prediction results by k-means clustering for topsoil and subsoil CEC were proximal ($\sigma_{p,c}^2$= 3.74 and 6.60, respectively), while the subsurface was best by the proximal plus remote sensed data ($\sigma_{p,c}^2$= 12.46). The clustering model for predicting soil pH, topsoil and subsoil were proximal ($\sigma_{p,c}^2$ = 0.33 and 1653.52, respectively), while the subsurface was best with proximal plus remote sensed data ($\sigma_{p,c}^2$ = 0.70). We conclude that linear and nonlinear regression models can be used for predicting soil CEC and pH in sandy infertile soil. However, the better performance was the LMM method.

KEYWORDS

sandy infertile soil, linear mixed model, fuzzy clustering, soil improvement guideline

1 | INTRODUCTION

Spatial predictions of soil properties are needed for various purposes including land use planning systems design for agriculture and engineering as well as scientific disciplines such as soil science, ecology, and hydrology [1]. For example, clay content maps can be used to determine land use potential [2], while soil pH maps can indicate lime requirement to counteract soil acidity or potential nutrient availability. However, the costs associated with soil sampling and laboratory analysis are substantial, and spatial prediction requires considerable sample effort given the observation by [3] that approximately 100 sample points are required to estimate a spatial statistical model. One way to improve soil sampling efficiency is to combine direct measurement of soil properties with the collection of remote data that are cheaper to measure. Remote data can be used to improve precision with which properties are predicted from relatively few direct observations. The growing interest in proximal geophysical sensing methods [4] has been applied to a range of problems including soil salinity assessment [5], prediction of depth to clay [6], soil moisture determination [4], determination of soil cation exchange capacity [7] and deep drainage estimation [8].

we consider two possible approaches. The first is to use remote data to form a set of land

classes using a numerical clustering algorithm. The mean value of the soil property in each class, estimated from samples within each class, can then be used for prediction. This approach could be useful because it makes no assumptions about the nature of the relationship between the soil property and the remote variables and because precise estimates of class means can be obtained from bulk samples formed by aggregating individual sample cores within the class, thus reducing analytical costs. One practical question for the implementation of this approach is how many classes should be defined. This is usually addressed by considering the distribution of the remote variables used to form the classes, looking for evidence of compact structures in feature space (see, e.g., [7]). The rationale of this approach is that the classes so-identified reflect natural clusters in the feature space rather than an arbitrary partition, and so should reflect underlying sources of variation in the soil. Another approach (not used in this context to date) is prediction-based. As we consider more and smaller classes, the within-class variance of the soil properties we wish to predict will, in general, diminish, but the prediction error does not necessarily because the class mean is estimated with less precision as a fixed sample effort is divided between more classes [9].

A second and more commonly used approach is linear predictive modeling, essentially a multiple regression of the target soil property on the remote variables. Ideally, this is done using data obtained from a probability sample so residuals can be treated as independent. The model is then used to form a prediction of the target property at a site where only remote data is known. Often data are not collected according to a probability design, in which case a linear mixed model (LMM) fitted in which covariates are fixed effects but the residuals are treated as a combination of a spatially correlated random effect and an independent and identically distributed error [10]. The prediction of the soil property at an unsampled site is then a combination of a regression-type prediction from proximally-sensed covariates and a kriging-type prediction of residuals from the fixed effects model at sampled sites (e.g. [11]).

K-means clustering algorithm is known as unsupervised classification technique in identifying similarities of particular objects based on vectorized distance [12]. Previously, some research papers in agricultural field were used for soil classification mapping [13]:[14]. There are increasing trends of using clustering techniques for proximally data [9].

In this paper we consider both approaches, showing how the question 'how many classes?' can be addressed in terms of the uncertainty of resulting predictions, and compared with the linear mixed model (LMM). We illustrate this with a case study in which handheld $\gamma - ray$ spectrometry and the apparent electrical conductivity using an electromagnetic (EM) induction instrument were measured as remote data across two fields located east of the village of Shelford near Nottingham in the UK. We formed classes from the remote data using k-means (KM) analysis. We then analyse data on soil properties along with the classes formed from the remote data and the remote data themselves. We show how the precision of class means as predictors of soil properties (for fixed total sample effort) varies with the number of classes and compare this criterion for the number of classes with measures based on the distribution of the remote data. We also compare these measures of precision with comparable ones for

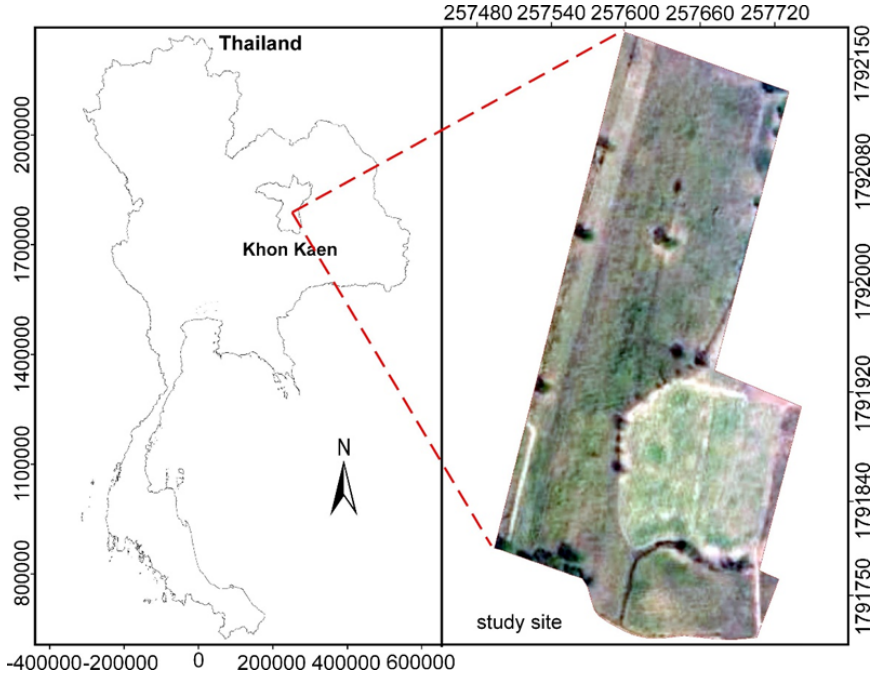

FIGURE 1 Study site in Khon Kaen Province, Thailand.

direct prediction from the remote data by a linear model.

2 | MATERIALS AND METHODS

2.1 | study area

The study site (Lat 16° 11' 41.404"N and Lon 102° 43' 54.504" E) is located in the Ban Haet district, Khon Kaen Thailand (Fig. 1), lying a short distance to the west of Ban Haet village, which is located approximately 40 km from Khon Kaen. The area is approximately 6 ha (154 m x 390 m) with the dominant soil type being an acid sandy loam to sandy Alfisols; described by Land Development Department of Thailand (scale of 1: 25,000). The current land use is rain-fed sugarcane farming system. The topography across the site is flat to relatively flat with a slope of 0-5 per cent. The climate is tropical savanna [15]. The mean annual precipitation is around 1,100 mm with the average minimum and maximum temperatures are 18.7 and 35.2^oC, respectively (TMD website, 2017). However, the area has 3 distinct seasons. The dry-season occurs between mid-February to mid-May with the hottest temperatures in April (43.9^oC) and minimum rainfall (224.4 mm). Conversely, the rainy season is between May to October, with an average temperature typified by July (24.4^oC) with some rainfall (1,104 mm). The winter season is between mid-October to mid-February, with minimum temperatures in December

(24.2^oC) with some rainfall (76.3 mm).

2.2 | Remote data, instruments and data collection

We collected four types of digital data. This included two proximal and two remote sensed data. The proximal data included $\gamma - ray$ spectrometry and electromagnetic (EM) induction. We collected $\gamma - ray$ data, from windows which measure three naturally occurring radioactive isotopes, including 40K (percentage - %), 232Th (parts per million - ppm) and 238U (ppm). In addition, the readings from this instrument were integrated to derive a measure of Total Count (ppm). We collected this data from a handheld RS-230 spectrometer (Radiation Solutions Inc. Mississauga, Ontario, Canada). It has a detector size of 103 cm^3. We collected the data from 254 sites on the 13 transects and on an approximate 15 x 15 m grid. The spectrometer was placed directly on the soil surface where it was bare otherwise plant cover (e.g. grasses or stubble) was removed before measurement. The assay scan length was set at 2 minutes. The field measurements were taken on 16 - 20 February 2018, which was in during summer and the soil surface was relatively dry.

We collected EM data using a EM38 [16] instrument. The instrument consists of a transmitter and receiver coil located at either end and spaced 1.0 m apart. It measures the apparent soil electrical conductivity (EC_a. The depth of exploration depends on coil configuration. In the horizontal mode, the EM38 measures EC_{ah} within a theoretical depth of 0-0.75 m. In the vertical mode, the EM38 measures EC_{av} within a theoretical depth of 0-1.5 m. The EC_a data was collected on the same 15 x 15 m grid described above. The EC_a data was collected on 16 February 2018. In all, 254 measurement sites were visited. All proximal sensed gamma-ray and EM38 EC_a data were georeferenced using a Garmin Etrex Legend G (Garmin International, Inc) submeter GPS.

2.3 | Remote sensed data acquisition

2.3.1 | Leaf area index (LAI)

LAI was derived from satellite Landsat 8 (OLI) image (30 x 30 m resolution) which acquired on 14thAugust 2017 by the United States of Geological Survey (USGS). The image had already done geometric correction from the source. The pre-processing method by ENVI 5.3 (Exelis Visual Information Solutions, Inc. 2015) was conducted using radiometric calibration, consequently the FLAASH atmospheric correction was applied to eliminate haze and moisture airdrop before extracting the digital number by ArcMap 10.4 (ESRI, 2017). The LAI was calculated using formula provided by [17]:

$$LAI = 0.57^{*exp(2.33*NDVI)}$$

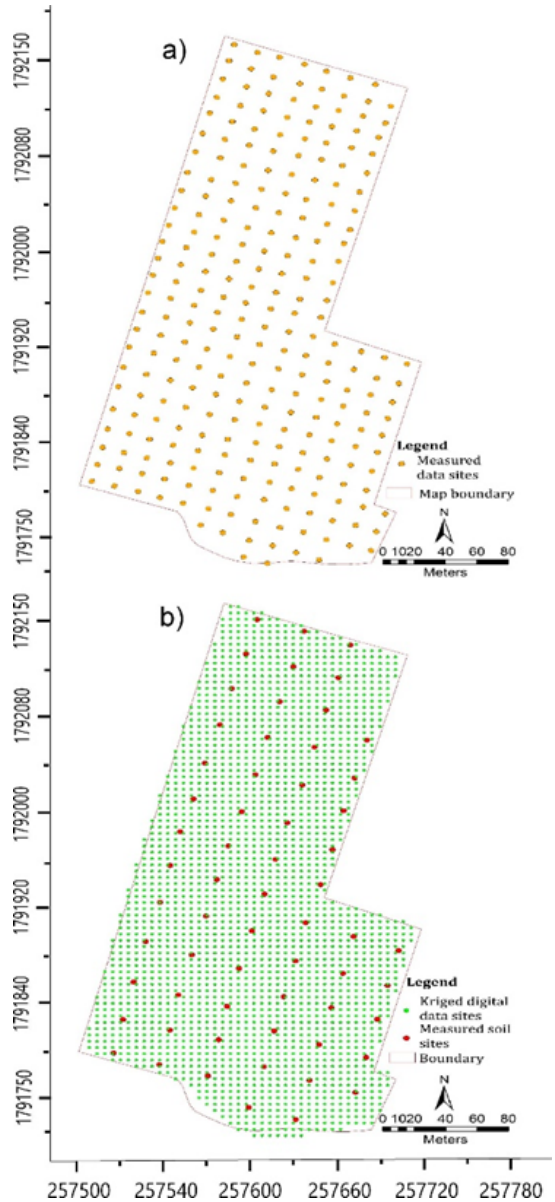

FIGURE 2 Map locations of; a) Measured and collected digital data (i.e. Gamma-ray, EM38hv (EC_{ah} and EC_{av}), LAI and DEM) sites, and b) Kriged digital data sites.

Where LAI denotes leaf area index and NDVI is normalized difference vegetation Index as,

$$= \frac{Nir - Red}{Nir + Red}$$

Where Nir denotes Near infrared band or band 5 and Red denotes the visible band 4 or red band of Landsat 8 (OLI) Regarding this, the 254 measurement sites were used as the input file and conversed to be a shape file for extracting the digital number on the LAI single layer image.

2.3.2 | Digital elevation model (DEM)

A five-metre resolution of DEM was acquired from Land Development Department (LDD), Thailand. The shape file of 254 measurement sites was used to extract elevation values from DEM in ArcGIS (ESRI, 2017). As the process, the extracted values were exported to be a text file for digital analysis.

2.4 | Soil Sampling and Laboratory Analysis

To determine if a direct linear relationship and non-linear relationship between various digital data could be developed the model with topsoil (0-0.3 m), subsurface (0.3-0.6 m) and subsoil (0.6-0.9 m) CEC and pH, The 61 sampling points were selected on the even number transect and every 30 metre across the study field in order to avoid the sampling bias. Fig. 2a shows the location of all sampling locations. The soil samples were air-dried, ground and passed through a 2-mm sieve. Laboratory analysis involved determination on the cation exchange capacity (CEC - cmol(+)$kg^{-}1$ based on ammonium saturation method. Regarding this, soil samples were saturated by NH_4OAc pH 7.0 and rinsed the $NH4^{+}$ using NaCl (Na^{+}). Then, the distillation apparatus and titrate method were used to determine CEC (Chapman,1965). Apart from this, soil pH was measured using a 1 part of soil per 1 parts of water dilution. The samples were collected on the 18 February 2018.

2.5 | Data analysis

The $\gamma - ray$ spectrometry (i.e. K, U, Th and TC), EM38 EC_{ah} and EC_{ah}, LAI and DEM data were first interpolated onto a common 5-m grid. This was done by ordinary kriging (OK) within a neighbourhood of 20-30 and a local variogram. The Vesper program was used [18]. Numerical clustering of the OK remote data was conducted by K-means clustering by JMP Pro13 software.

2.6 | Linear mixed model (LMM)

When soil data were collected according to a systematic sampling scheme, it would not be appropriate to fit a linear model by ordinary least squares, since the residuals cannot be treated as independent random variables. Rather we propose a linear mixed model for the data of the form;

$$y = X\beta + \eta + \varepsilon$$

where y is a n × 1 vector of values of the target soil variable, X is a n × p design matrix, β is a ρ × 1 vector of fixed effects coefficients, η is a n × 1 vector the elements of which are a realization of a spatially correlated random variable and ε, is a n × 1 vector the elements of which are a realization of an independent and identically distributed random variable. The elements of the design matrix are the predictor variables and the fixed effects coefficients correspond to these. For example, if the predictors are p classes then element {*i*,*j*} of X is 1 if the *ith* observation corresponds to the *jth* class and zero otherwise. There is therefore exactly one element equal to 1 in each row of the design matrix. In this case the elements of β are the estimated mean values of the target soil variable in the respective classes. The correlated random variable η is assumed to be normal and has mean zero and variance parameters which are familiar from the geostatistical literature. These are an overall variance, σ^2_η, and a distance parameter for a selected variogram function (e.g. the range of a spherical variogram). The error variable ε also has zero mean and a variance σ^2_ε. In an alternative form of this model the elements in the design matrix may be, in the first column, a column of ones and in the next p - 1 columns the values of the p - 1 proximally sensed auxiliary variable, in which case the vector β contains an intercept and p - 1 regression coefficients. We fitted models of the form in Eq. (7) for target soil variables and with the fixed effects either the class of maximum membership for the KM clustering of the remote variables with k = 2 or a subset of the remote variables in a regression type model. The fitting was done using the LME procedure from the NLME library for the R platform [19] ; [20]. Under this procedure, variance parameters for the random effects are first estimated by residual maximum likelihood (REML) and the fixed effects coefficients are then estimated by weighted least squares. The null hypothesis that all class means are equal (where the fixed effects are classes) or that the regression coefficients are zero (where the fixed effects are continuous variables), is tested by the Wald statistic. These methods are described by [10]. After model fitting summary statistics and histograms of the residuals were examined to check that these appeared consistent with an assumption of normality. The subset of remote variables for the model with continuous fixed effects was selected by first fitting a full model with all of the $\gamma - ray$ (K, U, Th and TC), EM38 EC_{ah} and EC_{av}, LAI and DEM data as predictors. This was then compared to a series of reduced models by dropping each predictor in turn, and the full and reduced models were compared by computing their log-likelihood ratio and testing this against chi-squared with one degree of freedom [21]. Any predictor where the

reduced model formed by dropping it was not significantly worse than the full model was rejected. This procedure was repeated until no further predictors were rejected. This procedure requires maximum likelihood rather than REML estimation since residual likelihoods cannot be compared between models with different fixed effects. Once a predictor set was selected the model was then re-estimated by REML.

2.7 | K-means (KM) clustering analysis

There are a number of research papers used various kinds of digital data such as airborne gamma-ray spectrometer survey data and electromagnetic (EM) induction survey data [9]: [22] to build up the soil properties classes using k-mean clustering algorithm. This includes the use of K-means algorithm to cluster EM with NDVI [23] or topographic witness index [24] or EM data with Quickbird imagery ([25] . In this research we use K-means clustering (Hard clustering) which described else where by [14] to cluster both proximal and also remote digital data.

The working by divides a data sample into exclusive clusters. The initial values of clusters' centroids are randomly selected from the availabl0e data. Updating centroids and clustering of data is then repeated until convergence is reached or for a defined number of iterations. A new centroid for a cluster is calculated based on each data sample that belongs to that cluster. The first issue of application of K-means-type algorithms is that the number of clusters should be known in advance.

K-means (KM) clustering is a widely used partitioning method. This method aims to make K mutually exclusive clusters of n data samples characterised by d parameters. Each cluster K is defined with one central point (centroid) determined by a certain combination of parameters contained in each data sample. KM is known as a method of vector quantization, since it is based on the location of points and their mutual distances. Namely, data samples described by d parameters can be presented as points in a d-dimensional space, where their coordinates are determined by the values of d parameters. A data sample belongs to a cluster defined with a centroid which is the closest one to the considered sample (point). The closest centroid is chosen after calculating the distance of each data from each centroid. Each data sample can belong to exactly one cluster. Hence, KM clustering is also called hard clustering.

The KM algorithm aims to distribute a set X of n data samples into K clusters. Each data sample is defined by d parameters. We consider data samples as points in a d dimensional space for better visualization. Input to the algorithm is the number of clusters K. The initial values of centroids $C_l^1 C_k^1 ... C_k^1$, $C_i \in R^d$, are chosen randomly from the available data samples. After the calculation of the distance of each data sample from set X to each clusters' centroid, each data sample is declared to be a member of its closest cluster. A set of data samples that belong to a cluster defined by centroid c_i is denoted as c_i, $I \leq i \leq K$. In each iteration, a centroid is estimated as a mean value of d corresponding parameters of all data samples which are a member of a corresponding cluster. Calculating K new centroids in each iteration is equivalent to changing a clusters' position in a d-dimensional space, till optimal m cluster

positions are reached. The processes of clustering and updating centroids are repeated until convergence has been reached or for a specified number of iterations. One drawback of KM clustering appears when a point is equally close to more than one centroid. In this case, the algorithm will not converge, since this point belonging will oscillate among a few different clusters, resulting in different clustering. However, this rarely happens in practice.

2.8 | Computation of the prediction error variance for regression models

For the case of prediction direct from a selected subset of proximally sensed variables by a multiple regression-type predictor the mean squared prediction error for a particular prediction at an unsampled site is:

$$\sigma_p^2 = \sigma^2(1 + xT\{X^T X\}^{-1}x)$$

where X is the design matrix for the data set used to predict the model and x is a vector in which the *ith* element is the difference between the value of the *ith* predictor for the particular prediction and the overall mean of the *ith* predictor [26]. The term σ^2 is the residual variance of the fitted regression. To compute the expected value of σ_p^2 for some simple random sample of size N we evaluated Eq. (2) for values of the selected predictor variables at each of the M = 254 observation sites and computed the average. We obtained the expression $X^T X$ by computing it from the design matrix, X, for our original M observations and rescaling it for a sample of size N and, as before, we used the sum of the REML estimates of the variances of the random effects in the corresponding LMM (Eq. (3)), σ_η^2,R + σ_ε^2,R, as the expected value of σ^2. The expected value of the mean squared prediction error from a regression estimated for some sample size N is therefore computed here as;

$$(\sigma_p^2(N) = (\sigma_{\eta,R}^2 + \sigma_{\varepsilon,R}^2) * 1 + \frac{1}{M\Sigma_{i=1},..Mx_i^T(\frac{N}{M})X^T X^{-1}x_i}$$

where xi is the vector of predictor values for the ith of our original M = 254 sampled locations and X is the design matrix for that same set. 2.9. Computation of the prediction error variance for class means The objective of this paper is to compare two approaches to the prediction of soil properties from the remote variables, given that a relatively small set of direct measurements of the soil property is available. To do this we compute the expected value of the mean squared prediction error for the alternative methods:

$$\sigma_{p,c}^2 = E[y - y'^2]$$

where y denotes the value of the target variable at some unsampled location and y' denotes the predicted value. When the predictor is the mean of a class (here obtained by cluster

analysis of the remote data) then the mean-squared prediction error in class i is

$$\sigma^2_{p,i} = \sigma^2_i(1 + \frac{1}{n_i})$$

where σ^2_i is the variance of the target property within class i and the mean of class i was estimated from *ni* independently and randomly selected observations within the class [27]. In this study we use a pooled within-class variance σ^2_w. If πi denotes the relative area of the *ith* class out of k and N is the total number of observations, then the expected value of the mean squared prediction error for classes is:

$$\sigma^2_{p,c} = \Sigma_{i=1},..k\sigma^2_w\pi_i(1 + 1 = N\pi I) = \sigma^2_w(1 + \frac{k}{N}).$$

In general, as k increase we expect the classes to become internally more uniform with respect to soil properties, so σ^2_w should decrease. However, it is apparently that the term in brackets on the right-hand side of Eq. (6) will increase with increasing k, and that this increase will be greater the smaller is N. In summary, $\sigma^2_{p,c}$ will only decrease with increasing k if the reduction in the within class variance is large enough to compensate for the fact that the fixed total sample size is spread more thinly over more classes which contributes to the uncertainty with which the class means are estimated. In this study we computed $\sigma^2_{p,c}$ for each target soil property for k = 2 classes formed by the KM algorithm. To do this we require a value of σ^2_w. This was obtained from the LMM, Eq. (7), fitted to the observed soil data for the corresponding classification. The sum of the variances of the random effects in the model for k classes as the random effects, $\sigma^2_{\eta,k} + \sigma^2_{\epsilon,k}$, was treated as the expected value of the variance for the random variable. This approach is used elsewhere to compute values for the variances of design-based sample estimates from the results of model-based analyses. [28] , [29] provides an example in soil science. The expected value of the mean squared prediction error for our classification into k classes for some sample size N is computed here as follows:

$$\sigma^2_{p,c}(\frac{N}{K}) = \sigma^2_{\eta,k} + \sigma^2_{\epsilon,k}(1 + \frac{k}{N})$$

3 | RESULTS AND DISCUSSION

3.1 | Contour plot of measured soil CEC and pH data

Fig. 3 shows the contour plot of measured CEC (cmol(+)kg^{-1}) and pH at different three depths including, topsoil, subsurface and subsoil from the 61 collected soil data in Table 1. Fig. 3a shows measured topsoil (0-0.3 m) CEC (cmol(+)kg^{-1}). There were two trends of CEC values

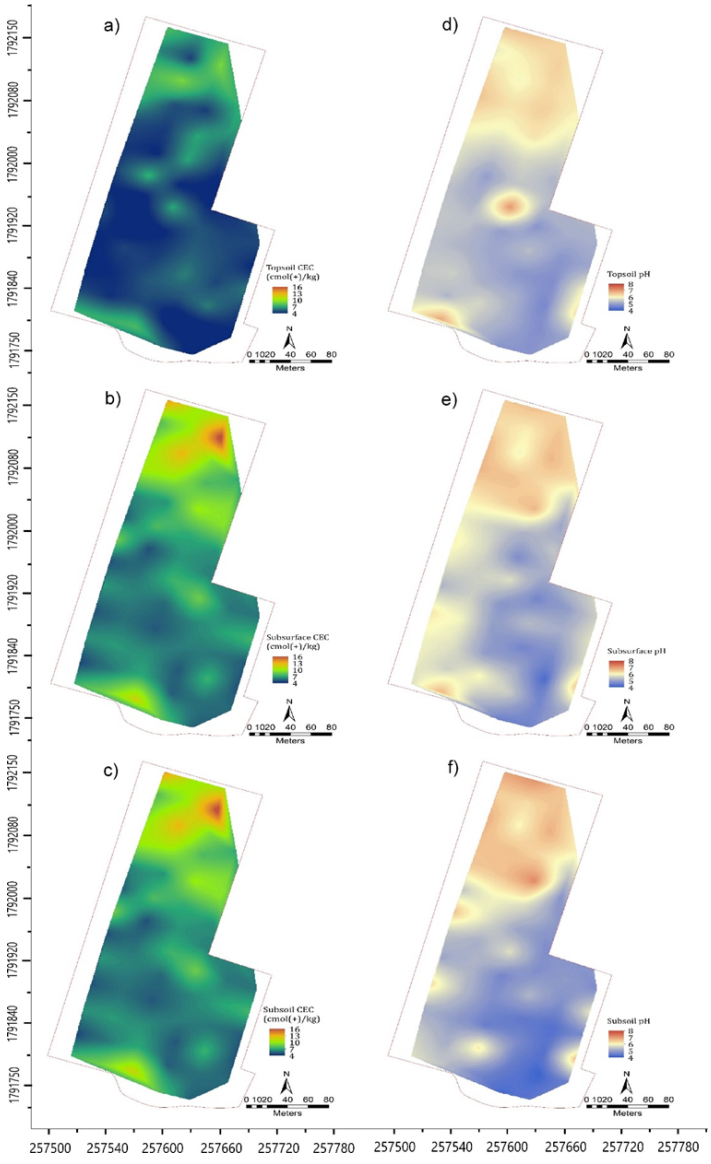

FIGURE 3 Contour plots of measured soil chemical properties includes; a) topsoil (0-0.3 m), b) subsurface (0.3-0.6 m) and c) subsoil (0.6-0.9 m) cation exchange capacity (CEC – cmol(+)kg^{-1}) and e) topsoil, f) subsurface and g) subsoil pH..

which the northern half of the study area was characterised by intermediate (5.0-7.0 cmol(+) kg^{-1}), including the two small parts in the middle area, southern left corner and around the east end of the southern half. While the majority of southern half was intermediate-small (4.0-5.0 cmol(+)kg^{-1}). Fig. 3b. shows the measured subsurface (0.3-0.6 m) CEC which generally was characterised by intermediate-small (4.0-5.0 cmol(+)kg^{-1}), especially in the southern half of the study area, exceptionally, the two small spots of the small (< 4.0 cmol(+)kg^{-1}) value in the middle half and in the southern right corner. The intermediate (5.0-7.0 cmol(+)kg^{-1}) to intermediate-large (7.0-10.0 cmol(+)kg^{-1}) was mostly in the northern half. Fig. 3c. shows the same trend to the measured subsurface CEC. The measured subsoil (0.6-0.9 m) CEC shows two main features by the intermediate (5.0-7.0 cmol(+)kg^{-1}) value in the northern half and within the west margin toward the southern end. While, the southern half was noticeably almost intermediate-small (4.0-5.0 cmol(+)kg^{-1}). This was consistently when compare to the topsoil and subsurface CEC.

Fig. 3d. shows distribution of measured topsoil (0-0.3 m) pH which was mainly characterised by the intermediate to intermediate-small (4.0 - 6.0) which mostly in the northern end toward the middle part of study area. This was not the case in some spots in the centre, southern west end and in the southeast margin where they were characterised by intermediate to intermediate-large (6.0-7.0). whereas the northern half was intermediate to intermediate-large (6.0-7.0). Fig. 2e. shows the measured subsurface (0.3-0.6 m) pH which was in the same trend to topsoil pH. The lager (> 6.0) pH was characterised in entire the northern half. Down to the middle toward the southern half, the measured soil pH was intermediate (6.0) to intermediate-small (5.0) in the west margin. Conversely, the small to intermediate-small (4.0-5.0) was in the east margin of the southern half. Fig. 2f. shows similar trend to the subsurface pH. The northern half toward the middle area was characterised by intermediate to large (> 6.0). Whereas, in the south-half was mainly characterised by small to intermediate small (4.0 – 5.0) and intermediate (6.0) in some spots.

3.2 | The contour plot of proximal and remote sensed data

Fig. 4 shows the contour plot of 61 proximal sensed data collected From Table 3 where soil data are available (Table 1.). using a RS-230 $\gamma - ray$ spectrometer (i.e. K, U, Th and TC) and EM38 EC_a data (i.e. EC_{ah} and EC_{av}). Fig. 4a shows the contour plot of K (%). It varies from intermediate-large (0.15-0.2 %) in the south but was for the most part intermediate (0.10-0.15 %). Fig. 4b shows the plot of Th (ppm). It varies from intermediate (4.5-5.5 ppm) in the north and some in the south but was for the most part intermediate-small (3.5-4.5 ppm). Fig. 4c shows the plot of Uranium (U-ppm). It was mainly characterised by intermediate (2.0-2.5 ppm). While the intermediate-large (2.5-3.0 ppm) to larger (3.0 ppm) was in around north right corner. Similarly, the fig. 4d shows the plot of total count (TC-cps) which was generally characterised by intermediate (24.0-36.0 cps.), while the small (< 12.0 cps.) was varies in between intermediate. Fig. 4e and 4f show the plot of measured EM38 ECa data (i.e. EM38 EC_{ah} and EM38

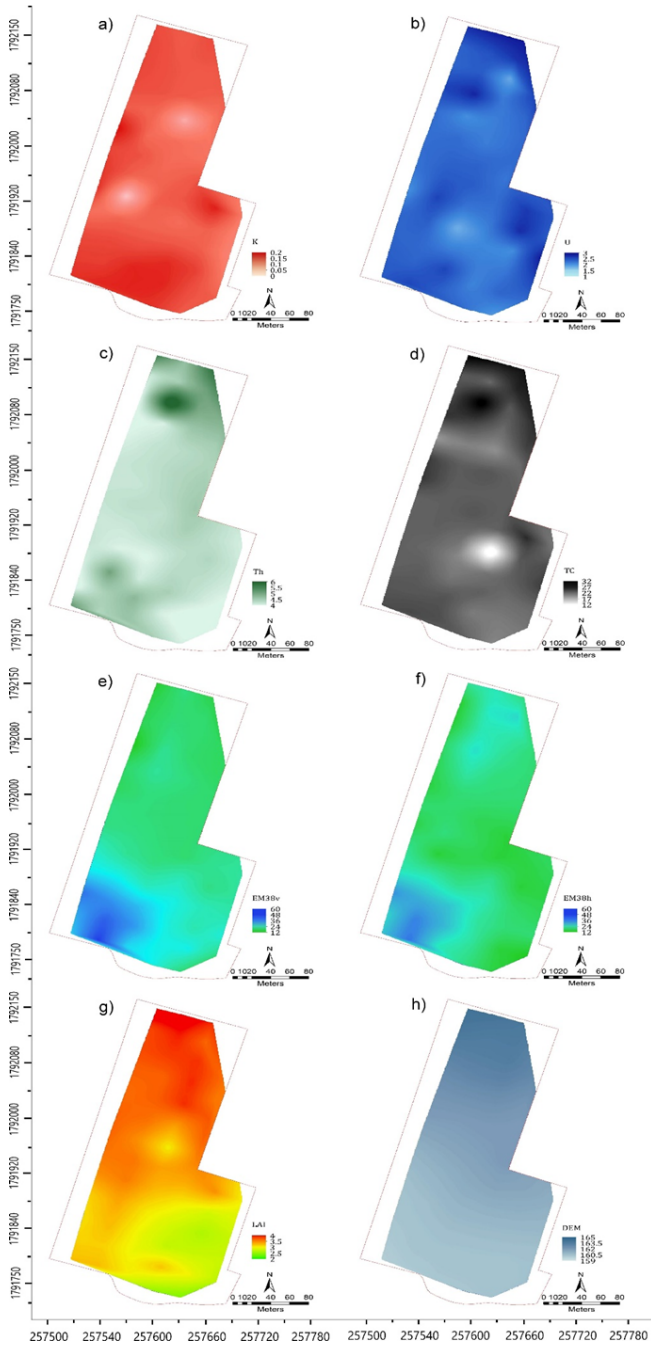

FIGURE 4 Contour plots of measured and collected digital data including; a) Uranium (^{238}U-ppm), b) Potassium (^{40}K-%), c) Thorium (^{232}Th-ppm), d) Total count (TC), e) Apparent electrical conductivity (EC_{av} - mSm^{-1}) from EM38v , f) Apparent electrical conductivity (EC_{av} - mSm^{-1}) from EM38h, g) Leave area index (LAI) and h) Digital elevation model (DEM)

TABLE 1 Summary statistic of 254 locations of proximally and remotelt survey data (e.g. γ-ray: K, U, Th and TC) and EM38 horizontal and vertical: EC_{ah} and EC_{av}) and remote sensed (Leaf area index: LAI and Digital Elevation Model: DEM) data.

Ancillary	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis	SE
K	0.13	0.03	0.13	0	0.25	0.25	-0.48	3.57	0
Th	4.39	0.47	4.29	3.14	6.55	3.41	1.42	3.31	0.03
U	2.29	0.27	2.25	1.33	3.4	2.07	0.64	2.31	0.02
TC	26.05	3.27	25.62	13.2	54.53	41.33	2.65	22.63	0.21
V0	27.83	9.31	24.57	0.04	56.27	56.23	0.95	0.57	0.57
H0	19.95	5.34	18.97	9.5	42.36	32.86	0.92	0.34	0.34
LAT	3.34	0.44	3.38	2.15	4.36	2.21	-0.17	-0.81	0.03
DEM	161.57	1.14	161.36	159.79	164.1	4.31	0.59	-0.86	0.07

EC_{av}) collected by EM38 instrument. Fig. 4e shows EC_{ah} data. It varies from intermediate (25.0-35.0 mSm^{-1}) in the south left corner. The majority part was intermediate-small (15-25.0 mSm^{-1}). Similar to the fig. 4e. but come up with a bigger range, fig. 4f. which shows the plot of EC_{av} data. It was characterised by intermediate-large (35.0-45.0 mSm^{-1}) in the south but was intermediate-small (15.0-25.0 mSm^{-1}) in the north. Fig. 4g shows the distribution of remote sensed data of Leaf area index (LAI) acquired from Landsat 8 (OLI) image and digital elevation model (DEM). The contour plot LAI data varies from intermediate-large (3.5-4.5) in the north and the intermediate (< 3.5) in the south. Similarly, Fig. 4f. shows the plot of DEM or digital elevation model. The elevation trend was decreasing from north to south. regarding this, the intermediate-large (163.0-165.0 m) was characterised in the north and intermediate (161.0-163.0 m) was in the middle, while the intermediate-small (159.0-161.0 m) was in the south.

3.3 | LMM analysis

We interpret the correlation between various potential predictor variables with small of mean prediction errors (σ_ρ^2) using linear mix model [30]. Table 3. shows the statistical correlation between soil properties at each depth with predictor variables from proximal, remote and proximal plus remote digital data from the LMM estimated by REML. The results show that, at the topsoil (0 - 0.3 m) CEC, it can be predicted by proximal data from U, Th, EC_{ah}, EC_{av} with suitable correlation (R^2) =0.57) and smallest error (0.001) and mean square prediction error (σ_ρ^2 = 1.931) when compare to Remote (σ_ρ^2 = 2.376) and Proximal plus Remote data (σ_ρ^2 = 1.980) respectively. Similarly, the subsurface (0.3-0.6 m) CEC can also be predicted by proximal data from K, TC, EC_{ah}, EC_{av} with a strong correlation (R^2 =0.74) and smallest error (0.03) and mean

TABLE 2 Summary statistic of 61 locations of proximally and remotely calibration data (e.g. γ-ray: K, U, Th and TC) and EM38 horizontal and vertical: EC_{ah} and EC_{av}) and remote sensed (Leaf area index: LAI and Digital Elevation Model: DEM) data.

Covariate	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis	SE
K	0.12	0.06	0.1	0.0	0.3	0.3	0.58	1.02	0.01
U	2.27	0.53	2.3	0.9	3.4	2.5	-0.17	0.05	0.07
Th	4.17	1.11	4.1	0.7	6.9	6.2	-0.04	0.68	0.14
TC	25.05	4.5	24.9	9.5	34.4	24.9	-0.57	0.31	0.58
H0	19.92	5.52	19.2	1.2	36.2	35.0	0.97	0.10	0.71
V0	28.07	9.67	25	3	58	55	1.05	1.09	1.24
LAT	3.35	0.44	3.3	2.39	4.24	1.85	-0.17	-0.79	0.06
DEM	161.57	1.14	161.28	159.97	164.1	4.14	0.64	-0.79	0.15

TABLE 3 Summary statistic of 61 locations of measured soil cation exchange capacity (CEC – cmol(+)kg^{-1}) and pH at different depths.

Soil_data	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis	SE
CEC (0-0.3 m)	5.39	1.69	4.88	2.81	10.03	7.22	0.78	-0.04	0.22
CEC (0.3-0.6 m)	7.49	2.39	6.71	4.81	16.71	11.9	1.52	2.56	0.31
CEC (0.6-0.9 m)	8.39	2.98	7.54	3.29	16.81	13.52	0.99	0.55	0.38
pH (0-0.3 m)	5.55	0.56	5.4	4.6	7.2	2.4	-0.44	-0.85	0.08
pH (0.3-0.6 m)	5.64	0.68	5.6	4.2	6.7	2.5	-0.06	0.25	0.09
pH (0.6-0.9 m)	5.48	0.87	5.4	4.1	7.2	3.1	0.28	-1.26	0.11

square prediction error (σ_ρ^2 = 2.599) when compare to Remote (σ_ρ^2 = 3.608) and Proximal plus Remote data (σ_ρ^2 = 2.715) respectively. This confirms by a strong correlation when compared 1:1 (LIN's = 0.72) between soil CEC versus predicted CEC. While the subsoil (0.6-0.9 m) CEC can only be predicted by Proximal plus Remote data from Y, K, EC_{ah}, EC_{av}, LAI and DEM with suitable correlation (R^2 =0.57) and smallest error (0.048) and mean square prediction error (σ_ρ^2 = 6.025) when compare to Remote (σ_ρ^2 = 6.638) and Proximal plus Remote data (σ_ρ^2 = 7.159) respectively.

In term of the soil pH prediction, Table 4 shows the topsoil (0 - 0.3 m) pH can be predicted by proximal plus Remote digital data from U, EC_{ah}, LAI, DEM with a strong correlation (R^2 =0.70) and smallest error (0.004) and mean square prediction error (σ_ρ^2 = 1.84). This was confirmed by a strong correlation when compared 1:1 (LIN's = 0.67) between soil pH versus predicted pH. This was almost the same with Proximal predictors that shows a strong correlation (R^2 = 0.68) and small mean prediction error (σ_ρ^2 = 0.185), followed by Proximal plus Remote data (σ_ρ^2

TABLE 4 Statistical analysis of Digital Soil Mapping using Linear mix models (LMM) of proximal [P] (Potassium: ^{40}K, Uranium: ^{238}U, Thorium: ^{232}Th and Total count: TC (cps), EM38 Horizontal mode: EC_{ah}, and EM38 Vertical mode: EC_{av}) and remotely [R] sensed (Leaf area index: LAI and Digital Elevation Model: DEM) data for cation exchange capacity (CEC – cmol(+)kg^{-1}).

Digital data	Predictors	Corr	ME	RMSE	LIN'S	σ_ρ^2
P (0-0.3 m)	U, Th, EC_{ah}, EC_{av}	0.57	0.001	1.389	0.52	1.931
R (0-0.3 m)	LAI, DEM	0.48	-0.001	1.479	0.40	2.376
PR (0-0.3 m)	U, Th, EC_{ah}, EC_{av}, LAI	0.55	-0.000	1.407	0.51	1.980
P (0.3-0.6 m)	K, TC, EC_{ah}, EC_{av}	0.74	0.030	1.612	0.72	2.599
R (0.3-0.6 m)	LAI, DEM	0.60	0.014	1.899	0.55	3.608
PR (0.3-0.6 m)	TC, EC_{ah}, DEM	0.72	0.025	1.647	0.70	2.715
P (0.6-0.9 m)	K, EC_{ah}, EC_{av}	0.50	0.022	2.576	0.44	6.638
R (0.6-0.9 m)	LAI, DEM	0.44	0.029	2.675	0.37	7.159
PR (0.6-0.9 m)	Y, K, EC_{ah}, EC_{av}, LAI, DEM	0.57	0.048	2.460	0.53	6.052

Note: Y = Northing, Corr = correlation, σ_ρ^2 = mean square prediction error, ME = mean error, LIN's = lin's concordance coefficient and RMSE = root mean square error

= 0.217) respectively. While the subsurface (0.3-0.6 m) and subsoil (0.6-0.9 m) soil pH were best predicted by Proximal data. The Y-coordinates, U and EC_{av} predicted subsurface soil pH with a strong correlation (R^2 =0.69) small error (0.008) and mean square prediction error (σ_ρ^2 = 0.184). The subsoil pH was best predicted by U and EC_{av} with a highest correlation (R^2 = 0.76), small error (0.003) and small prediction error (σ_ρ^2 = 0.312) when compare to others predictors, however they all are still in a strong correlation (R^2 > 0.7) and small prediction error (σ_ρ^2 = 0.350 and 0.368, respectively).

3.4 | KM analysis

We divide various digital data samples into several clusters using K-mean hard clustering method. First, we assigned number of clusters from K = 2 into all member digital data. Theoretically, the unsupervised machine learning (K-means: KM) algorithm can help grouping each proximal and remote digital data points (254 points) into clusters and based on their inherent properties and calculate their similarity using the Euclidean distance as a measurement. We compare these clusters with soil properties at each depth from topsoil (0-0.3 m) subsurface (0.3-0.6 m) and subsoil (0.6-0.9 m) CEC and pH. The results showed different trends in all three depths as Table 5. To predict the topsoil (0-0.3 m) CEC, the results show all variables digital data from both proximal, remote and their combinations (proximal plus remote) provided a similar trend in all statistical values. However, the best optimal cluster was created by a remote variable with the highest nugget (0.467) smallest range (354.383) and clustering prediction error ($\sigma_{\rho,c}^2$ = 3.735).

TABLE 5 Summary statistic for analysis of Digital Soil Mapping using Linear mix models (LMM) of proximal [P] (Potassium: ^{40}K, Uranium: ^{238}U, Thorium: ^{232}Th and Total count: TC (cps), EM38 Horizontal mode: EC_{ah}, and EM38 Vertical mode: EC_{av}) and remotely [R] sensed (Leaf area index: LAI and Digital Elevation Model: DEM) data for soil pH.

Digital data	Predictors	Corr	ME	RMSE	LIN's	σ_{ρ}^2
P (0-0.3 m)	Y, U, EC_{ah}	0.68	0.005	0.430	0.67	0.185
R (0-0.3 m)	LAI, DEM	0.62	0.018	0.466	0.57	0.217
PR (0-0.3 m)	U, EC_{ah}, LAI, DEM	0.70	0.004	0.428	0.67	0.184
P (0.3-0.6 m)	Y, U, EC_{av}	0.69	0.008	0.490	0.66	0.184
R (0.3-0.6 m)	LAI, DEM	0.60	0.002	0.542	0.55	0.294
PR (0.3-0.6 m)	U, EC_{av}, LAI	0.66	0.010	0.508	0.63	0.259
P (0.6-0.9 m)	U, EC_{av}	0.76	0.003	0.558	0.74	0.312
R (0.6-0.9 m)	LAI, DEM	0.71	0.002	0.606	0.68	0.368
PR (0.6-0.9 m)	EC_{ah}, LAI	0.73	-0.000	0.592	0.70	0.350

Note: Y = Northing, Corr = correlation, σ_{ρ}^2 = mean square prediction error, ME = mean error, LIN's = lin's concordance coefficient and RMSE = root mean square error

The best predictor for subsurface (0.3-0.6 m) CEC was from the proximal data. Whereas the subsoil (0.6-0.9 m) CEC was the combinations digital data. Table 6 shows the suitable clustering predictor for soil pH at the three depths. In the topsoil (0-0.3 m) pH, the most suitable cluster from proximal data was provided the largest nugget (0.348) and with smallest prediction error ($\sigma_{\rho,c}^2$ = 0.326). while the remote digital data clustering was suitable for subsurface (0.3-0.6 m) soil pH with large nugget (0.253), small range (209.853) and low prediction error ($\sigma_{\rho,c}^2$ = 0.497) when compared to other predictors. While the clustering results from all digital data cannot be provided an optimal result when compared to collected subsoil (0.6-0.9 m) pH.

3.5 | Spatial distribution of predicted CEC and pH by LMM

Fig. 5 shows the distribution of predicted CEC (cmol(+)kg^{-1}) and pH at different three depths including, topsoil, subsurface and subsoil. Fig. 5a shows predicted topsoil (0-0.3 m) CEC (cmol(+)kg^{-1}). There were three major trends of CEC values. The small (4-7 cmol(+)kg^{-1}) to intermediate small (7-10 cmol(+)kg^{-1}) is the cast in the south-east corner. Whereas the intermediate small to intermediate (7-10 cmol(+)kg^{-1}) CEC was in the southwest corner. This is not the case in the north corner where the CEC is generally shown around intermediate (10 cmol(+)kg^{-1}). Similar but in different degree to the subsurface (0.3-0.6 m) and subsoil (0.6-0.9 m) CEC, fig. 5b and 5c shows the same trends, the predicted subsurface CEC was generally characterised by small to intermediate-small (4-7 cmol(+)kg^{-1}) in the southwest corner, while in the southern was mostly intermediate small (7 cmol(+)kg^{-1}) to intermediate-

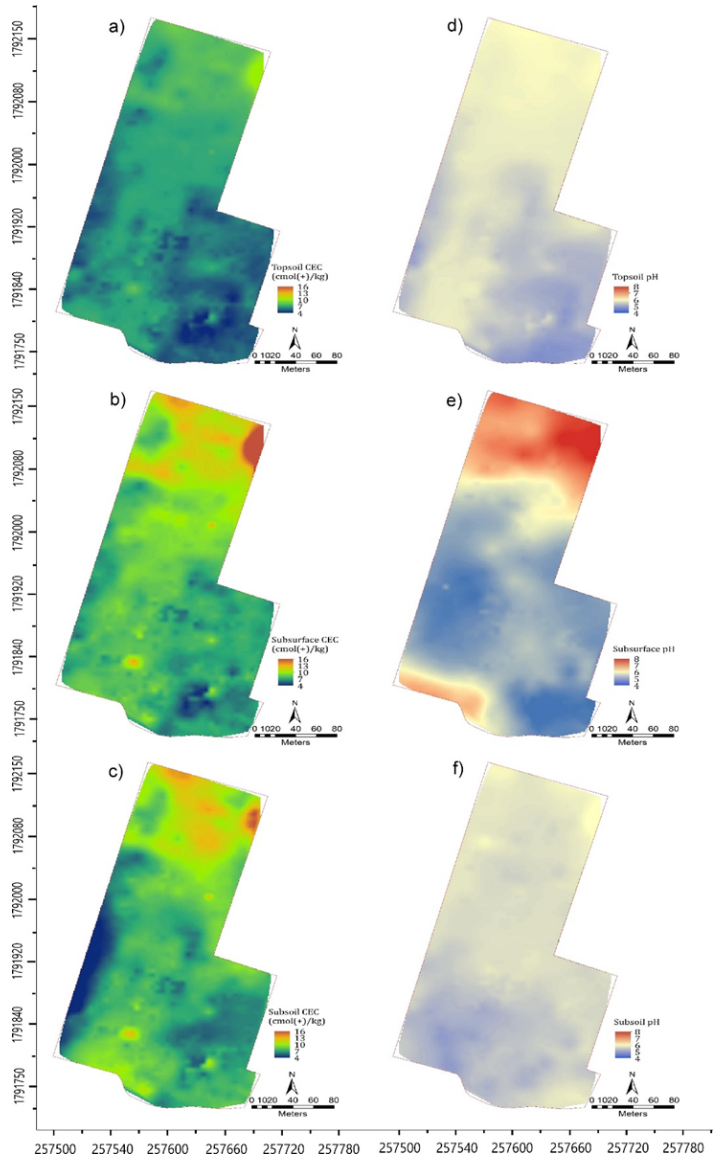

FIGURE 5 Spatial distribution of predicted cation exchange capacity (CEC – cmol(+)kg^{-1}) at depths of a) topsoil (0-0.3 m) derived from Proximal data (U, Th, EC_{ah}, EC_{av}), b) subsurface (0.3-0.6 m) from Proximal data (K, TC, EC_{ah}, EC_{av}) and c) subsoil (0.6-0.9 m) from Remote data (Y, K, EC_{ah}, EC_{av}, LAI, DEM) and soil pH at d) topsoil (0-0.3 m) derived from Proximal and Remote data (U, EC_{ah}, LAI, DEM), e) subsurface (0.3-0.6 m) from Proximal data (Y, U, EC_{av}) and f) subsoil (0.6-0.9 m) from (U, EC_{av}) generated using linear mixed model (LMM)

TABLE 6 Statistical analysis of Digital Soil Mapping using clustering of proximal [P] and remote [R] digital data for cation exchange capacity (CEC – cmol(+)kg^{-1}).

Digital data	Cluster	Nugget	Range	SD	$\sigma^2_{\rho,c}$
P (0-0.3 m)	2	0.437	355.185	1.946	3.912
R (0-0.3 m)	2	0.467	345.383	1.902	3.735
PR (0-0.3 m)	2	0.450	355.060	1.926	3.833
P (0.3-0.6 m)	2	0.372	361.285	2.529	6.604
R (0.3-0.6 m)	2	0.267	532.009	2.946	8.969
PR (0.3-0.6 m)	2	0.363	361.971	2.550	6.716
P (0.6-0.9 m)	2	0.085	125.059	3.534	12.903
R (0.6-0.9 m)	2	0.048	128.702	3.656	13.812
PR (0.6-0.9 m)	2	0.105	124.638	3.473	12.460

Note: SD = Standard deviation and $\sigma^2_{\rho,c}$ = mean square prediction error

large (10 cmol(+)kg^{-1}). while the northern conner was mostly characterised by intermediate (10 cmol(+)kg^{-1}) to intermediate large (13 cmol(+)kg^{-1}) and large in the northeast corner (16 cmol(+)kg^{-1}). Fig. 5c. shows the topsoil CEC varies from small to intermediate small (4-7 cmol(+)kg^{-1}) in the south west corner and intermediate small to intermediate (7-10 cmol(+)kg^{-1}) in the middle part. While, in the south west corner was small (4 cmol(+)kg^{-1}). The northern corner was shown the same trend to the subsurface CEC and was around intermediate (10 cmol(+)kg^{-1}) to intermediate large (13 cmol(+)kg^{-1}). The only spot of northeast corner was large (16 cmol(+)kg^{-1}).

Fig. 5d shows the distribution of predicted pH of topsoil (0-0.3 m). It was the same trends of CEC, with an extremely acid (4.0) in the southeast corner, while the southwest corner was characterised by extremely to very strongly acid (4.0-5.0). The middle-half up to the northern corner was clearly characterised by moderately acid (6.0). This was not the case for the subsurface soil pH (Fig. 5e). The distribution was characterised by extremely acid (4.0) in the southeast corner up to entire of the middle part of the field. The slightly alkali (7.0) to moderately alkali (8.0) was characterised by the small part southwest edge and the northern corner. Fig. 5f shows the same distribution of the topsoil pH with very strongly acid (5.0) to moderately acid (6.0) in the southern part of the field, while the northern part was characterised by moderately acid to neutral (6.0-7.0).

3.6 | Comparison of mean squared prediction error of the map

In order to determine whether classifying the proximal and remote digital data using the K-means clustering algorithm compare to linear mixed model (LMM). The LMM result provides

better prediction of soil properties than the clustering model, as we compared the $\sigma^2_{p,c}$ and σ^2_p , respectively. First, we developed the regression models for CEC and pH. The model parameters calculated from the R platform are shown in Tables 3 and 4. In terms of CEC mapping in the topsoil (0-0.3 m), the combination of proximal data was the best predictor with aa small mean square prediction error (σ^2_p = 1.931). This is not the case in the pH of the top soil (0-0.3 m), where the smallest mean square prediction error (σ^2_p = 0.184) was from the combination of proximal and remote data. However, the CEC of the subsurface and subsoil provided by a combination of proximal and proximal plus remote data, respectively, created a greater mean square prediction error (σ^2_p = 2.599 and 6.052 respectively) than those of the pH of the soil at the same depths (σ^2_p = 0.184 and 0.312 respectively).

4 | CONCLUSION

It can be seen that spatial predictions of soil properties, such as CEC and soil pH, can be acquired using various reliable methods. This is because all the soil properties are within the scorpan framework which includes: soil (s), climate (c), organisms (o), relief (r), parent material (p), age (a), and spatial location (n) that can influence the variation of soil properties. Therefore, it can be fitted quantitative relationships between soil properties or classes. These include linear mixed models, classification and regression trees, neural networks, fuzzy systems and geo-statistics [31]. Regarding this, all data used in this research can be described as they are reliable to the environmental factors. The handheld gamma-rays (Uranium, Potassium, Thorium and Total count) can be linked to mineral properties and soil parent materials, while the EM38 EC_{av} and EC_{ah} provide direct relationship with soil moisture, clay content and salinity. The leaf area index (LAI) provides organism constraint and digital elevation model means to relief of the study sit. The linear mixed model (LMM) and fuzzy clustering were applied to determine if all digital data can be predicted soil CEC and pH. We examine all digital data and all soil properties using LMM first and then K-mean clustering, to see if there is a linear or non-linear relationship among those interested data. Fortunately, there were both linear and non-linear relationships from only CEC and pH can be found. However, small prediction error (σ^2_p < 0.5) can be found only for soil CEC in all depths. While the soil pH can only provide good prediction in the topsoil and subsurface. This is might be because of the high sandy textural property affects soil pH [32]. Therefore, we conclude that the better performance was the LMM.

Regarding this result, it was working on small study area (< 10 ha). This is the reason for the small different variation as we seen by predicted soil properties distribution. However, to create the large areas of soil properties, the approach can be used and begins with these two models.

acknowledgements

This research was fully financial supported by Land Development Regional Office 5, Land development Department (LDD), Ministry of Agriculture and Cooperatives, Thailand. Special thanks for The University of New South Wales on Supporting research facilitation during that time. Finally, this research can be finished on 12th Friday, September 2025 with good guidance from Prof. John Triantafilis who dedicated his life keep teaching me with valuable research principles.

references

[1] Gooverts P. Geostatistics for natural resources evaluation: Oxford University Press. New York 1997;.

[2] Khongnawang T, Zare E, Zhao D, Srihabun P, Triantafilis J. Three-dimensional mapping of clay and cation exchange capacity of sandy and infertile soil using EM38 and inversion software. Sensors 2019;19(18):3936.

[3] Webster R, Oliver MA. Sample adequately to estimate variograms of soil properties. Journal of soil science 1992;43(1):177–192.

[4] Robinson DA, Abdu H, Lebron I, Jones SB. Imaging of hill-slope soil moisture wetting patterns in a semi-arid oak savanna catchment using time-lapse electromagnetic induction. Journal of Hydrology 2012;416:39–49.

[5] Lesch S, Corwin D, Robinson D. Apparent soil electrical conductivity mapping as an agricultural management tool in arid zone soils. Computers and Electronics in Agriculture 2005;46(1-3):351–378.

[6] Jung WK, Kitchen NR, Sudduth KA, Anderson SH. Spatial characteristics of claypan soil properties in an agricultural field. Soil Science Society of America Journal 2006 7;70(4):1387–1397. [Online; accessed 2025-09-10].

[7] Triantafilis J, Lesch SM, La Lau K, Buchanan SM. Field level digital soil mapping of cation exchange capacity using electromagnetic induction and a hierarchical spatial regression model. Soil Research 2009;47(7):651–663.

[8] Woodforth A, Triantafilis J, Cupitt J, Malik R, Subasinghe R, Ahmed M, et al. Mapping estimated deep drainage in the lower Namoi Valley using a chloride mass balance model and EM34 data. Geophysics 2012;77(4):WB245–WB256.

[9] Huang J, Davies GB, Bowd D, Monteiro Santos FA, Triantafilis J. Spatial prediction of the exchangeable sodium percentage at multiple depths using electromagnetic inversion modelling. Soil Use and Management 2014 feb 26;30(2):241–250. [Online; accessed 2025-09-10].

[10] Lark R, Cullis B, Welham S. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. European Journal of Soil Science 2006;57(6):787–799.

[11] Gooley L, Huang J, Pagé D, Triantafilis J. Digital soil mapping of available water content using proximal and remotely sensed data. Soil use and management 2014;30(1):139–151.

[12] Valsalan P, Sriramakrishnan P, Sridhar S, Latha GCP, Priya A, Ramkumar S, et al. Knowledge based fuzzy c-means method for rapid brain tissues segmentation of magnetic resonance imaging scans with CUDA enabled GPU machine. Journal of Ambient Intelligence and Humanized Computing 2020;p. 1–14.

[13] McBratney A, De Gruijter J. A continuum approach to soil classification by modified fuzzy k-means with extragrades. Journal of Soil Science 1992;43(1):159–175.

[14] Vukčević M, Popović-Bugarin V, Dervić E. DBSCAN and CLARA Clustering Algorithms and their usage for the Soil Data Clustering. In: 2019 8th Mediterranean Conference on Embedded Computing (MECO) IEEE; 2019. p. 1–6.

[15] Yoothong K, Moncharoen L, Vijarnson P, Eswaran H. Clay mineralogy of Thai soils. Applied Clay Science 1997;11(5-6):357–371.

[16] McNeill JD. Geonics Limited Technical Note TN-35 2012;.

[17] Fang H, Baret F, Plummer S, Schaepman-Strub G. An overview of global leaf area index (LAI): Methods, products, validation, and applications. Reviews of Geophysics 2019;57(3):739–799.

[18] Budiman Minasny A, McBratne B, Whelan BM. Vesper-program for automatic Variogram Estimation and Spatial Prediction with ERror. Pedometrics, 99 1999;p. 66.

[19] Pinheiro J, Bates D, DebRoy S, Sarkar D, Team R, Linear and nonlinear mixed effects models in R; 2013.

[20] Core R. Development Team. R: Version 2010;3(2).

[21] Molenberghs G, Verbeke G. Meaningful statistical model formulations for repeated measures. Statistica Sinica 2004;p. 989–1020.

[22] Zare E, Huang J, Santos FM, Triantafilis J. Mapping salinity in three dimensions using a DUALEM-421 and electromagnetic inversion software. Soil Science Society of America Journal 2015;79(6):1729–1740.

[23] Dang TH, Ngo LT, Pedrycz W. Multiple kernel based collaborative fuzzy clustering algorithm. In: Asian Conference on Intelligent Information and Database Systems Springer; 2016. p. 585–594.

[24] Priori S, Barbetti R, L'Abate G, Bucelli P, Storchi P, Costantini EA. Natural terroir units, Siena province, Tuscany. Journal of Maps 2014;10(3):466–477.

[25] Guo Y, Shi Z, Li H, Triantafilis J. Application of digital soil mapping methods for identifying salinity management classes based on a study on coastal central China. Soil Use and Management 2013;29(3):445–456.

[26] Dudewicz EJ, Mishra S. Modern mathematical statistics. John Wiley & Sons, Inc.; 1988.

[27] Brus D, Lark R, Soil surveys. In 'Encyclopedia of environmetrics'. Wiley Online (John Wiley & Sons Inc.: Hoboken, NJ, USA); 2013.

[28] Cochran RN, Horne FH. Statistically weighted principal component analysis of rapid scanning wavelength kinetics experiments. Analytical Chemistry 1977;49(6):846–853.

[29] Lark RM. Spatially nested sampling schemes for spatial variance components: scope for their optimization. Computers & Geosciences 2011;37(10):1633–1641.

[30] Huang J, Scudiero E, Choo H, Corwin DL, Triantafilis J. Mapping soil moisture across an irrigated field using electromagnetic conductivity imaging. Agricultural Water Management 2016 1;163:285–294. [Online; accessed 2025-09-10].

[31] McBratney AB, Santos MM, Minasny B. On digital soil mapping. Geoderma 2003;117(1-2):3–52.

[32] Hamarashid NH, Othman MA, Hussain MAH. Effects of soil texture on chemical compositions, microbial populations and carbon mineralization in soil. Egypt J Exp Biol(Bot) 2010;6(1):59–64.