Sustainable Environment Research

RESEARCH Open Access

Estimating topsoil texture fractions by digital soil mapping - a response to the long outdated soil map in the Philippines

Jeremy P. Mondejar[1,2*] and Alejandro F. Tongco[1]

Abstract

Digital soil mapping for soil texture is mostly an understanding of how soil texture fractions vary in space as influenced by environmental variables mainly derived from the digital elevation model (DEM). In this study, topsoil texture models were generated and evaluated by multiple linear regression (MLR), ordinary kriging (OK), simple kriging (SK) and universal kriging (UK) using free and open-source R, System for Automated Geoscientific Analyses, and QGIS software. Comparing these models is the main objective of the study. The study site covers an area of 124 km^2 of the Municipality of Barili, Cebu. A total of 177 soil samples were gathered and analyzed from irregular sample points. DEM derivatives and remote sensing data (Landsat 8) were used as environmental variables. Exploratory analyses revealed no outlier in the data. Skewness and kurtosis values of the untransformed data vary greatly between −3.85 to 7.20 and 1.8 to 70.7, respectively; an indication that variables are highly skewed with heavy tails. Thus, Tukey's ladder of powers transformation was applied that resulted to normal or nearly normal distribution having skewness values close to zero and kurtosis values have lighter tails. All data analysis from MLR modeling, variography, kriging, and cross-validations of models were implemented using the transformed data. Forward selection, backward elimination, and stepwise selection methods were adapted for predictors selection in MLR. The MLR, OK, SK, and UK were applied and cross validated for topsoil texture prediction. Likewise, exponential, Gaussian, and spherical models were fitted for the experimental variograms. Backward elimination method for clay, sand, and silt have the lowest MAE and highest R^2 in MLR. The UK fitted with exponential variogram model has the highest R^2 of 0.878, 0.821, and 0.893 for clay, sand, and silt, respectively. These models can be adapted as a decision support for agricultural land use planning and crop suitability development in the area.

Keywords: Geostatistics, Mountain ridge proximity, Multiple linear regression, Ordinary kriging, Simple kriging, Universal kriging

Introduction

Dealing with global and regional challenges in land degradation, food security, water scarcity, and climate change, an accurate and updated geospatial soil information is imperatively needed [1, 2]. These problems are directly related to soil functions particularly to agricultural productivity, loss of biodiversity, and provision of water [3]. Producing accurate and reliable soil maps is indispensable in watershed management [4, 5], rangeland management, and landscape ecology [6].

The traditional soil survey process is tediously difficult to update rapidly and accurately. This process has associated significant limitations. First, significant changes in environmental conditions are not readily observed, especially when processing several variables simultaneously; secondly, the entire process must be repeated for each update that makes soil survey updates very inefficient [7]. Conventional soil survey adapts the manual process of producing a polygon-based soil map, whereby, without the computer-based approach, the map cannot be

* Correspondence: mondejar.jeremy13@gmail.com
[1]School of Engineering, University of San Carlos – Talamban Campus, Cebu City 6000, Philippines
[2]Department of Agricultural and Biosystems Engineering, Cebu Technological University – Barili Campus, Barili 6036, Philippines

updated rapidly and accurately as the entire production procedure must be repeated [7]. Such method is time-consuming, requires numerous soil samples, and expensive [8]. Geographic information systems (GIS) can overcome this problem with the application of digital soil mapping (DSM). The DSM estimates soil properties by establishing interrelationships between soil properties and the environmental variables derived mainly from the digital elevation model (DEM) and remotely sensed images [8, 9]. Thus, the direction of DSM is toward the generation of dynamic and replicable geospatial soil information [10].

GIS algorithms have been adapted for an efficient spatial interpolation technique in land resources inventories [11, 12] in addressing the limitation of the traditional soil survey. GIS is a tool for data input, handling, analyzing and output process. It plays a significant role in spatial decision-making that involves information collection for DSM. GIS can perform several tasks using both spatial and attribute data and can integrate a variety of geographic technologies like Global Positioning System (GPS) and Remote Sensing (RS). GIS integrates spatial and geostatistical analysis, and the efficient management, storage, and retrieval of geographic data [8, 9, 13]. Thus, GIS as a tool plays a significant role in the implementation of a computer-based spatial decision making support system.

The slow progress in agricultural production and the steadily increasing population in the Philippines require the applications of this computer-based decision support system. The Philippine Department of Agriculture is still using over 40-year-old soil information in its programs for climate change mitigation and land use plan [14]. Bureau of Soils and Water Management is still in traditional soil survey method and is yet to implement DSM [15]. With these enormous challenges, there is an urgent necessity that soil maps in agricultural areas in the country be updated applying the modern technology of DSM.

Like most of the parametric test, multiple linear regression (MLR) requires normally distributed and homoscedastic variables [16, 17]. It is helpful for data analysis and inferences that both dependent and independent variables that are highly skewed [17, 18], and where standard deviations among variables significantly differ [16] should be transformed to normal or nearly normal distribution [19]. Thus, in this study, all variables were transformed to meet the conditions for normal or nearly normal distribution, minimum error and unbiased estimate [20].

Hence, this study aimed in establishing relationships between topsoil texture fractions (clay, sand, and silt) and the environmental variables by applying and comparing MLR, ordinary kriging (OK), simple kriging (SK) and universal kriging (UK) using transformed data integrating the use of free and open source software (FOSS) of R, System for Automated Geoscientific analyses (SAGA) GIS, and QGIS.

This study presents a successful approach in DSM using FOSS that are of best cost advantage to be adapted by any GIS users from a developing country like the Philippines. The achieved methodology can lead to valuable outcome in achieving a more comprehensive land use plan since the generated results are useful for watershed management particularly for ecological, hydrological, and crop suitability modeling.

Materials and methods

Materials

Synthetic Aperture Radar (SAR) DEM was acquired from MacDonald, Dettwiler and Associates, British Columbia, Canada and post-processed by the UP Training Center for Applied Geodesy and Photogrammetry, through the DOST-GIA funded Disaster Risk and Exposure Assessment for Mitigation Program which is downloadable through https://lipad.dream.upd.edu.ph/. This SAR DEM has 10 m resolution and a projection of WGS84 UTM Zone 51. A cloud-free Landsat 8 image in the study area was selected and downloaded from the United States Geological Survey data archive (https://earthexplorer.usgs.gov) for RS data. The study makes use of an adequate hardware setup with 32 GB multi-core processor, 64-bit operating system, and solid-state drive to avoid hanging or crashing [21] necessary for GIS and RS processing.

The study area

Barili, the study area, is a second income class municipality, with an annual income of ₱45 M (0.87 M USD) or more but less than ₱55 M (1.06 M USD), in the southern part of Cebu Province, Philippines. The municipality has a population of 73,862 based on 2015 census. Barili is located 60 km southwest of Cebu City (Fig. 1). It has an area of about 124 km^2 (about 9.5 × 13 km) with mostly agricultural lands. The elevation ranges from sea level to 540 m. It is bordered to the northeast with Carcar City and Sibonga; to the southwest with Dumanjug and the west lies the Tañon Strait; and to the northwest with Aloguinsan. Mountain ridgeline separates the watersheds of Barili in the west and Carcar City and Sibonga in the east.

Soil sampling, preparation, and analysis

A total of 177 soil samples were collected within the Municipality of Barili, Cebu, Philippines. A pit measuring approximately 1 × 1 × 1 m was dug manually in each physiographic position to examine and sample the soil profile. About 2 kg of composite soil sample were obtained from every horizon of each soil profile. Soil sampling was done randomly as shown in Fig. 1. Soil samples were air dried, freed of rocks and plant materials, ground using a wooden mallet and allowed to pass through a 2-mm sieve. Soil texture analyses were done at the Regional Soils

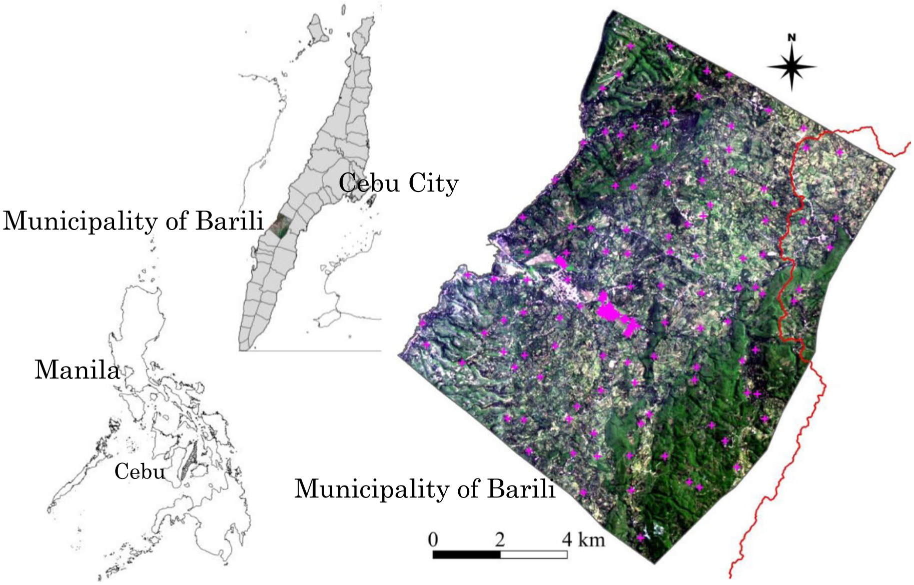

Fig. 1 The location of the Municipality of Barili (natural color composite) as the study area in the Province of Cebu. The distribution of samples (magenta), and the mountain ridgeline (red)

Laboratory, Department of Agriculture, Regional Field Office VII, Cebu City, Cebu, Philippines.

Environmental predictors

A total of 37 parameters were generated and analyzed to understand their relationship with soil texture properties. Environmental variables or predictors were the preprocessed sink-filled or depressionless DEM and its derivatives, geographic coordinates, and RS data as shown in Table 1. Majority of the predictors were generated from the preprocessed sink-filled DEM using SAGA GIS modules for basic terrain analysis, terrain analysis–channels, terrain analysis–hydrology, terrain analysis–morphometry, and terrain analysis–lighting. The predictors include (i) 7 parameters by basic terrain analysis module such as analytical hillshading, aspect, channel network base level, slope length and steepness (LS) factor, profile curvature, slope, and valley depth; (ii) 4 parameters by terrain analysis–channels such as overland flow distance, horizontal overland flow distance, vertical overland flow distance, and vertical distance to channel network; (iii) 3 parameters by terrain analysis–hydrology such as catchment area, flow accumulation, and SAGA wetness index; (iv) 7 parameters by terrain analysis–morphometry such as curvature classification, mass balance index, mid slope position, multiresolution index of valley bottom flatness (MRVBF), normalized height, slope height, and terrain ruggedness index; (v) 1 parameter by terrain analysis–lighting such as sky view factor. Furthermore, slope steepness (S) factor for universal soil loss equation (USLE) was generated applying *r.watershed* module of geographic resources analysis support system in QGIS. Mountain ridge proximity and stream proximity were generated using *proximity grid* module of SAGA GIS. Location of samples by their latitude and longitude, determined by GPS, were also included as predictors.

Remotely sensed image of operational land imager (OLI) and thermal infrared sensor (TIRS) instruments of Landsat 8 on the study area dated July 27, 2017 were also considered as additional predictors and was preprocessed using the semi-automatic classification plugin (SCP) using QGIS. The remotely sensed predictors include 7 bands (bands 1 to 7) of OLI and 2 bands (band 10 and 11) of TIRS.

Preprocessing of remotely sensed data

The SCP for QGIS was applied for Landsat 8 conversion to top of atmospheric reflectance and brightness temperature, and pansharpening [27]. Multispectral image analysis requires conversion of its “quantized and calibrated scaled digital numbers” [23] to top of atmosphere reflectance in order to achieve clear Landsat scenes [28] which is packaged in SCP. Preprocessing of Landsat 8 image was discussed in details by Congedo [27] and in the study by Mondejar and Tongco [29].

Exploratory data analysis

The exploratory data analysis consisted of three steps and was performed in R with contributing packages ‘rcompanion’ for plotting histogram and performing

Table 1 Environmental variables, description, descriptive statistics, and references

Parameters	Description	Minimum	Median	Mean	Maximum	CV (%)	Skewness	Kurtosis	References
Dependent Variables									
Clay, %	< 0.002 mm particle size	18.8	47.6	47.9	76.2	27.5	0.16	2.19	
Sand, %	0.05–2 mm particle size	4.4	20.4	22.6	65.4	52.7	0.81	3.51	
Silt, %	0.002–0.05 mm particle size	10	30	29.4	46.6	25.4	−0.22	2.56	
Independent Variables or Predictors									
Analytical hillshading, radians	The angle between the surface and the incoming light beams	0.14	0.797	0.83	1.86	37.6	0.35	3.53	[22]
Aspect, radians	Grid cells of aspect or facing direction	0.12	3.28	3.23	6.28	56.2	−0.04	1.79	[22]
Catchment area, m^2	Upslope area of each grid cell contributing to runoff	100	845.8	1982.6	39,526.3	192	6.50	58.3	[22]
Channel network base level, m	Interpolated channel network base level elevations	4.62	68.03	109.9	539.7	109	1.66	5.84	[4, 8]
Curvature classification	Surface curvature based terrain classification	0.000	3.70	3.699	8.00	65.5	0.14	1.77	[4]
Elevation, m	Digital Elevation Model, 10 × 10 m	4.4	68.2	110.3	540.2	109	1.66	5.85	[4, 22]
Flow accumulation	Number of upslope cells	100	879	2053	46,247	206	7.20	70.7	[22]
Landsat 8 Band 1	Coastal/Aerosol	14.81	16.18	16.09	18.22	3.1	0.43	5.25	[4, 5, 23]
Landsat 8 Band 10	Thermal Infra-Red Sensor 1	10.81	11.99	12.00	14.10	3.9	1.09	6.47	[4, 5, 23]
Landsat 8 Band 11	Thermal Infra-Red Sensor 2	0.032	0.045	0.045	0.076	1.6	1.13	5.95	[4, 5, 23]
Landsat 8 Band 2	Blue band	0.049	0.067	0.067	0.105	14.1	0.80	4.71	[4, 5, 23]
Landsat 8 Band 3	Green band	0.039	0.061	0.061	0.101	18.7	0.71	3.93	[4, 5, 23]
Landsat 8 Band 4	Red band	0.241	0.336	0.335	0.508	13.5	0.42	3.80	[4, 5, 23]
Landsat 8 Band 5	Near Infra-Red band	0.119	0.169	0.172	0.309	15.4	1.02	6.01	[4, 5, 23]
Landsat 8 Band 6	Short Wave Infra-Red 1 band	0.056	0.084	0.086	0.161	17.5	1.18	6.34	[4, 5, 23]
Landsat 8 Band 7	Short Wave Infra-Red 2 band	0.142	0.797	0.829	1.861	37.6	0.35	3.53	[4, 5, 23]
Latitude, m	UTM Northing: the y coordinate from the equator	1,110,588	1,117,281	1,117,805	1,125,070	0.2	0.33	3.27	
Longitude, m	UTM Easting: the x coordinate from the central meridian	551,858	557,678	558,067	564,098	0.4	0.20	3.26	
LS factor	Slope length and steepness factor as used by the Universal Soil Loss Equation (USLE)	0.030	0.181	0.573	10.814	224.7	4.96	32.14	[22, 24]
Mass balance index	Characteristics of soil and accumulation	−0.561	−0.007	−0.008	0.626	−3609	0.14	2.16	[4]
Mountain ridge proximity, m	Distance to the mountain ridge	207.700	4772	4434	10,473	48.4	0.10	3.14	This study
MRVBF	Flatness on valley bottom	0.000	0.000	1.271	5.000	129.3	1.03	2.63	[4, 25]
Mid-slope position	Relative vertical distance to the mid-slope valley or crest directions	0.012	0.675	0.595	0.964	41.9	−0.60	2.21	[4, 25]

Table 1 Environmental variables, description, descriptive statistics, and references *(Continued)*

Parameters	Description	Minimum	Median	Mean	Maximum	CV (%)	Skewness	Kurtosis	References
NDVI	Normalized difference vegetation index	0.440	0.702	0.691	0.807	9.2	−0.87	3.76	[4]
Normalized height	Height ratio above the channel to the ridge	0.018	0.240	0.352	0.973	81.7	0.75	2.14	[4]
Overland flow distance, m	Overland flow distance to the drainage channel	6.174	31.587	36.9	101.161	60.0	0.76	2.87	[13]
Profile curvature, m^{-1}	Shape of the surface in the immediate neighborhood	−0.010	0.000	0.000	0.012	−2204	0.34	3.81	[13, 26]
S factor	Slope steepness factor as used by the Universal Soil Loss Equation (USLE)	0.030	0.165	0.314	3.283	145.5	3.26	17.11	[24]
SAGA wetness index	Topographic Wetness Index (TWI) based on a modified catchment area	5.446	8.811	8.825	13.079	20.5	0.20	2.12	[22, 25]
Sky view factor	Ratio of visible sky viewed from the ground	0.881	0.994	0.989	1.000	1.7	−3.85	20.03	[4]
Slope, %	Slope gradient	0.000	4.943	6.067	28.510	85.7	1.54	5.87	[4, 22]
Slope height, m	Vertical distance from the base of the slope to the crest	1.023	4.582	9.655	86.743	125.1	2.65	13.20	[4, 25]
Stream proximity, m	Proximity to stream	3.706	278	608	3218	117.1	1.47	4.75	[4, 8]
Terrain ruggedness index	Terrain heterogeneity	0.000	0.636	0.784	3.845	84.7	1.72	6.83	[22]
Valley depth, m	Vertical distance to a channel network base level	0.774	11.244	18.669	177.309	131.8	3.86	21.28	[22]
Vertical distance to channel network, m	Altitude above the channel network in the same units as the elevation data	0.000	0.132	0.263	1.355	119.7	1.42	4.22	[8, 22]
Vertical overland flow distance, m	Vertical distance projected of mean runoff length	0.083	2.252	2.954	14.593	90.4	1.89	7.17	[4]

Tukey's ladder of power transformations; 'moments' for determining skewness and kurtosis; 'nortest' for determining different normality tests. Step 1: statistical data distribution was checked by plotting its histogram. Step 2: determination of skewness and kurtosis before data transformation. And Step 3: perform Tukey's ladder of power transformation.

Transformation of data

Normal distribution of data is an assumption in geostatistical analyses similar to several statistical techniques [16, 17, 30, 31]. In instances of non-normality, especially for strongly skewed data, data transformation to normality or at least symmetric distribution is needed [30, 31]. Normality test of the data can be implemented by simple evaluation of its skewness (close to 0) and kurtosis (close to the range of 1 to 3), normal q-q plot, and inferential test for normality such as Shapiro-Wilk or Kolmogorov-Smirnov tests [32].

Soil texture fractions and all of the predictors were transformed applying Tukey's ladder of powers transformation in R to ensure normality or near normality of dependent and independent variables. In this study, skewness and kurtosis evaluation were applied. The Tukey's ladder of powers transformation is defined as:

$$y = \left\{ \begin{array}{c} x^{\lambda}\,\mathrm{if}\,\lambda > 0 \\ logx\,\mathrm{if}\,\lambda = 0 \\ -\left(x^{\lambda}\right)\mathrm{if}\lambda < 0 \end{array} \right\} \tag{1}$$

where λ is the power coefficient parameter in transforming values of a parameter to follow a normal distribution as close as possible [17, 33]. In the rcompanion package, Shapiro-Wilk tests were performed iteratively and determine lambda value "that maximizes the W statistic" [17]. Furthermore, loop function in R was adapted to determine a particular transformation that had a distribution with skewness as close to zero and kurtosis as close to the range between 1 to 3.

Predictive models

The procedures tested to predict soil properties (sand, silt, and clay) were MLR, OK, SK and UK. The geostatistical procedures were implemented in SAGA GIS software [34]. MLR has been widely used to predict the response of a dependent variable from a set of independent variables, as a function of the correlations between them. The MLR analyses were executed in SAGA GIS applying forward selection, backward elimination, and stepwise selection methods, fitting the model by identifying variables which have most significance according to 95% confidence level.

MLR for prediction

Multiple regression seeks to determine the equation that best predicts the dependent variable Y as a linear function of a set of independent variables X where the linear relationship is expressed as:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3.... + \epsilon \tag{2}$$

where $\hat{Y}$ is the estimated value of Y for a given set of X values; a is the intercept; b_1 is the estimated slope (partial regression coefficient) of a regression of Y on X_1, considering the rest of X variables to be held constant, likewise for b_2, b_3, and so on; and ϵ is the error term [16]. In this study, the coefficient of multiple determination (R^2) is considered in the goodness-of-fit of a linear model for cross validation of models determined by the following equation as:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \dot{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \dot{y}\right)^2} \tag{3}$$

where $\hat{y}_i$ as the predicted or fitted values of the dependent variable, $\dot{y}$ is the mean of the observed values of the dependent variable, y_i as the observed values of the dependent variable, and n is the number of observed values of the data set.

Relative importance of environmental predictors

Relative importance is the "quantification of an individual regressor's contribution to a multiple regression model" [35]. In this study, relative importance metrics are forced to percentages or summed to 100%. Gromping [35] calls this "lmg" as proposed by Lindeman et al. [36]. This is achieved using package "relaimpo" in R by the following code: calc.relimp (linmod, type = c("lmg"), rela = TRUE).

Geostatistical analyses

Geostatistical analyses in this study were done making use of SAGA GIS 7 adapting the spatial and geostatistics-kriging module for grid interpolation from irregular sample points with variogram dialog. DSM is one of the many useful applications in SAGA GIS for soil science [34]. Variogram models were first generated before kriging techniques for the predictive models of soil fractions.

Semivariogram Kriging is based on the idea of spatial correlation or dependence that the degree of influence of observed points to unobserved points is inversely related to distance [20, 37, 38]. The first essential step in kriging is the calculation of empirical variogram before any geostatistical interpolation. In this study, isotropic semivariogram is determined considering only on the effect of h [39]. A semivariogram model estimates the relationship between semivariances or the differences between neighboring values and separation distance [40], such as the spherical or exponential models. The empirical semivariogram formula is as follows:

$$\gamma(h) = \frac{1}{2N(h)}\sum_{i=1}^{N(h)}\left[Z(x_i+h) - Z(x_i)\right]^2 \tag{4}$$

where $\gamma(h)$ is the semivariance of the lag distance between two sample points; $N(h)$ is the number of observation pairs of random variables separated by distance h; and $Z(x_i)$ and $Z(x_i + h)$ are the values of variables located at x_i and $x_i + h$, respectively [39, 41, 42]. The empirical semivariogram for each soil texture fraction was fitted with theoretical semivariogram models in SAGA GIS using its variogram module. In this study, the spherical, exponential and Gaussian models were fitted into the empirical variograms for the dependent soil fraction variables (clay, sand, silt).

Spherical:

$$\gamma(h) = 0 \text{ if } h = 0 \tag{5}$$

$$\gamma(h) = n + (s-n)\left[\frac{3h}{2r} - \frac{1}{2}\left(\frac{h}{r}\right)^3\right] \text{ if } 0 < h < r \tag{6}$$

$$\gamma(h) = n + (s-n) \text{ if } h > r \tag{7}$$

Exponential:

$$\gamma(h) = n + (s-n)\left[1 - exp\left(-\frac{h}{r}\right)\right] \tag{8}$$

Gaussian:

$$\gamma(h) = n + (s-n)\left[1 - exp\left(-\left(\frac{h}{r}\right)^2\right)\right] \tag{9}$$

where n is the nugget effect at zero distance h, s is the sill, and r is the range [39, 43–45].

SK Generally, weighted average of observed data points is an approach of spatial interpolation that is expressed as:

$$\gamma(h) = \frac{1}{2N(h)}\sum_{i=1}^{N(h)}\left[Z(x_i+h) - Z(x_i)\right]^2 \tag{10}$$

The empirical semivariogram for each soil texture fraction was fitted with theoretical semivariogram models in SAGA GIS using its variogram module. In this study, the spherical, exponential and Gaussian models were fitted into the empirical variograms for the dependent soil fraction variables (clay, sand, silt). The approximation of SK is based on the equation:

$$\hat{z}_{sk}(x_o) = \sum_{i=1}^{n}\lambda_i z(x_i) + \left[1 - \sum_{i=1}^{n}\lambda_i\right]\mu \tag{11}$$

where μ is a given stationary mean of the observed values that is assumed to remain constant throughout the domain, λ_i is the kriging weight assigned at sampled locations [31, 46, 47] wherein the kriging weights are estimated by minimizing the variance [46].

OK The OK is analogous to SK but OK involves that the summation of weights equals to one, such that $[1 - \sum_{i=1}^{n}\lambda_i] = 0$, with an accompanying Lagrange parameter ψ. Additionally, μ is not constant but it is recalculated within the search window across the modeled area of interest [46, 47]. Eventually, OK is estimated as:

$$\hat{z}_{ok}(x_o) = \sum_{i=1}^{n}\lambda_i z(x_i) \tag{12}$$

In order to obtain unbiased estimations for OK, the following equations are solved simultaneously:

$$\sum_{j=1}^{n}\lambda_j\gamma(x_i, x_j) + \psi = \gamma(x_i, x_0) \tag{13}$$

$$\sum_{j=1}^{n}\lambda_j = 1 \tag{14}$$

where $\gamma(x_i, x_j)$ is the value of the variogram between two points x_i and x_j, ψ and is the Lagrange parameter for the minimization of kriging variance [31, 37, 41].

UK OK and SK assume stationary process considering the mean. However, in other cases, spatial data are not stationary in the mean [31, 48] which is being considered in UK. Geostatistical literatures have used the terms UK, kriging with external drift, and regression kriging that are essentially adapting similar techniques [40, 49]. In UK, prediction at unvisited location is estimated by combining the predicted drift and residuals as:

$$\hat{z}(x_0) = u(x_0) + \hat{e}(x_0) \tag{15}$$

where drift or trend $u(x_0)$, a linear regression mean, is fitted by linear regression analysis, and $\hat{e}(x_0)$ is the interpolated residuals [40, 44, 49]. The trend can be expressed as a functional form:

$$u(x) = \sum_{k=0}^{K}\beta_k f_k(x) \tag{16}$$

where β_k, $k = 0, 1, \ldots, K$ are unknown coefficients while $f_k(x)$ are known functions of x [31, 48]. In this study, the best model resulted from MLR analyses for each soil fraction was applied and predictions at unobserved locations are estimated by the following equation:

$$\hat{z}_{uk}(x_o) = \sum\nolimits_{i=1}^{n} \lambda_i \left(\sum\nolimits_{k=0}^{K} \beta_k f_k(x_i) + \hat{e}(x_i) \right) \tag{17}$$

Cross validation

Cross validation technique was adopted for evaluating and comparing the performance of models. The models were evaluated applying k-fold cross validation due to the absence of validation data set and small size of observed samples for model validation [50, 51]. In this study, 5-fold cross validation was adapted. There is no strict rule on the choice of k, but 5 or 10 is usually applied [51]. Cross validation of models was done in R applying caret package. The mean absolute error (MAE), root mean squared error (RMSE), and R^2 were determined to evaluate the accuracy of models [8, 52, 53].

$$MAE = \frac{1}{n} \sum\nolimits_{i=n}^{1} |y_i - \hat{y}_i| \tag{18}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{19}$$

where y_i is observed value, $\hat{y}_i$ is the predicted value, and n is the number of samples.

Additionally, MLR models require homoscedasticity and normality of residuals [6, 54]. Thus, a scatter plot of residuals and predicted values were visually evaluated for homoscedasticity using *plot* function where a horizontal line at zero was added using *abline* function in R. Additionally, its normality test (skewness, kurtosis, Lilliefors test) along with its quantile-quantile (QQ) plot was determined in R applying plot diagnostics of *plot* function. Hence, homoscedasticity and normality of residuals were critical validation criteria as being meet by each model in this study.

Results and discussion

Exploratory data analyses

In this study, geostatistical analysis of data started with summary statistics, data transformation (Table 2), MLR modeling, variography, kriging, and mapping. The exploratory analysis revealed no outlier in the soil texture fractions (clay, sand, silt) and environmental variables. Thus, all 177 samples were used for the spatial interpolation. A statistical summary of the untransformed environmental variables (Table 1) shows high coefficient of variation between −3609 to 224.7%. Skewness and kurtosis values vary greatly between −3.85 to 7.20 and 1.77 to 70.7, respectively. Indicating that the variables are highly skewed with heavy tails, thus, do not fit the normality assumption. The Tukey's ladder of powers transformation was applied in transforming the data that resulted in normal or nearly normal distribution having skewness values close to zero and kurtosis values have thinner tails (Table 2). Consequently, all data analyses from MLR modeling, variography, kriging, and cross validations of models were implemented using the transformed data.

MLR modeling

The MLR analyzed the relationships between the transformed topsoil texture fractions and the transformed environmental variables. Table 3 shows the MLR results where forward selection, backward elimination, and stepwise selection methods were applied and cross-validated. At this stage, MLR established correlation between transformed dependent and independent variables. The dependent variables were the transformed values of soil texture fractions and the independent variables were the transformed values of environmental predictors. All predictors selection methods in MLR, in this study, revealed Lilliefors test values greater than 0.05. These were indications that the transformation of values both for dependent and independent variables resulted in an improvement of normality and variance stability [19], minimum error and unbiased estimate [20]. The predictors of selection methods with most homogeneity of variance (homoscedasticity) in terms of residuals were adapted as predictor grids, respectively, for clay, sand and silt fractions for UK as implemented in SAGA. The cross validation of MLR models are shown in Table 3 and their relative importance will be subsequently discussed.

Cross validation of MLR models

Prior to kriging, evaluation of regression models was implemented. Table 3 shows the cross validation of models applying different regression selection methods and the normality statistics of residuals. The MAE, RMSE, homoscedasticity and normality (skewness, kurtosis and Lilliefors test) test of residuals were considered in assessing the accuracy performance of linear predictive models. Homoscedasticity and normality test of residuals were applied in order to verify and satisfy the assumption of homogeneity of variance [54] towards choosing the best linear unbiased estimator. All results of the regression models were coupled with kriging for the implementation of UK to further evaluate their performance and be able to determine the best unbiased models. Stepwise selection method has the advantage of having the least number of significant predictors, while forward selection method had the lesser number of significant predictor compared to backward elimination method.

Based on error metrics, performance of forward and stepwise selection methods for clay fraction were the same because they revealed the same significant predictors; this was also observed for sand fraction. Backward

Table 2 Descriptive statistics of transformed variables and the corresponding transformation

Parameters	Tukey's power transformation	Minimum	Median	Mean	Maximum	CV (%)	Skewness	Kurtosis
Dependent Variables								
Clay	(x - 5)^(0.375)	6.54	8.46	8.44	10.01	9.23	0.000	2.20
Sand	(x)^(0.3)	1.56	2.47	2.47	3.51	16.52	−0.009	2.39
Silt	(63 - x)^(0.5)	4.05	5.75	5.76	7.28	11.25	−0.054	2.67
Independent Variables or Predictors								
Analytical hillshading	(x + 2)^(0.075)	1.06	1.08	1.08	1.11	0.82	0.002	3.30
Aspect	x	0.12	3.28	3.23	6.28	56.45	−0.041	1.79
Catchment area	-(x - 12)^(−0.075)	−0.71	−0.60	−0.60	−0.45	−9.13	0.005	2.47
Channel network base level	(x + 1)^(0.125)	1.24	1.70	1.68	2.20	15.01	0.175	1.71
Curvature classification	(23 - x)^(1.6)	76.16	113.98	114.87	150.93	19.82	−0.058	1.75
Elevation	$\log_{10}$(x)	0.64	1.83	1.76	2.73	30.02	0.024	1.60
Flow accumulation	(x - 15)^(0.075)	1.40	1.66	1.65	2.238	11.38	0.207	2.34
Landsat 8 Band 1	-(−x)^(−1)	12.54	23.61	24.00	31.39	11.17	0.108	5.01
Landsat 8 Band 10	(x - 13)^(0.375)	1.25	1.54	1.52	1.86	6.06	−0.093	4.30
Landsat 8 Band 11	(x - 10)^(0.125)	0.97	1.09	1.09	1.19	2.89	−0.018	4.53
Landsat 8 Band 2	-(x)^(−0.8)	−15.75	−11.98	−12.13	−7.82	−11.63	−0.016	3.40
Landsat 8 Band 3	-(x)^(−0.55)	−5.28	−4.42	−4.45	−3.46	−7.48	0.006	3.14
Landsat 8 Band 4	-(x)^(−0.375)	−3.37	−2.86	−2.89	−2.36	−6.83	−0.029	2.69
Landsat 8 Band 5	-(x + 1)^(−2.475)	−0.59	−0.49	−0.49	−0.36	−8.27	0.008	3.11
Landsat 8 Band 6	-(x + 1)^(−10)	−0.32	−0.21	−0.21	−0.07	−20.91	0.022	2.95
Landsat 8 Band 7	-(x)^(−0.725)	−8.06	−6.01	−6.05	−3.76	−11.85	−0.037	3.34
Latitude	$\log_{10}$(x)	6.05	6.05	6.05	6.05	0.02	0.319	3.27
Longitude	$\log_{10}$(x)	5.74	5.75	5.75	5.75	0.03	0.190	3.26
LS factor	$\log_{10}$(x)	−1.52	−0.74	−0.81	1.03	−86.99	0.398	1.95
Mass balance index	(3.4 - x)^(1.525)	4.74	6.48	6.51	8.16	12.97	−0.066	2.12
Mountain ridge proximity	(x)^(0.95)	159.10	3124.60	2895.20	6592.80	46.60	0.016	3.08
MRVBF	(x - 3)^(3)	−27.00	−27.00	−14.58	8.00	−89.63	0.253	1.36
Mid-slope position	-(x - 3)^(−4)	−0.06	−0.03	−0.03	−0.01	−36.45	0.005	1.81
NDVI	-(x - 1)^(−1)	1.79	3.36	3.36	5.18	19.03	0.062	2.48
Normalized height	-(x + 2)^(−7.575)	0.00	0.00	0.00	0.00	−62.97	−0.004	1.54
Overland flow distance	(x + 10)^(0.075)	1.23	1.32	1.32	1.42	3.59	0.000	2.14
Profile curvature	-(x + 1)^(−10)	−1.11	−1.00	−1.00	−0.89	−3.68	0.176	3.64
S factor	$\log_{10}$(x)	−1.52	−0.78	−0.89	0.52	−68.47	0.208	1.57
SAGA wetness index	(x - 2)^(0.475)	1.80	2.49	2.47	3.13	12.80	−0.026	2.06
Sky view factor	-(1 - x)^(−0.125)	−2.72	−1.91	−1.88	−1.31	−11.82	0.011	3.82
Slope	(x)^(0.475)	0.00	2.14	2.12	4.91	45.98	0.098	3.11
Slope height	(x - 1)^(0.075)	0.75	1.10	1.09	1.40	12.23	−0.048	2.06
Stream proximity	(x + 6)^(0.125)	1.33	2.03	2.03	2.75	18.65	−0.036	1.83
Terrain ruggedness index	-(x + 1)^(−0.85)	−1.00	−0.66	−0.66	−0.26	−25.45	0.038	2.47
Valley depth	-(x + 3)^(−0.325)	−0.65	−0.42	−0.42	−0.18	−22.37	0.017	2.89
Vertical distance to channel network	(x)^(0.3)	0.00	0.55	0.54	1.10	52.97	−0.075	2.10
Vertical overland flow distance	(x)^(0.175)	0.65	1.15	1.14	1.60	16.23	−0.043	2.81

Table 3 Cross validation of models by MLR applying different predictors selection methods

Texture fraction	Selection method	Predictors	MAE	RMSE	R^2	Residuals Normality statistics		
						Skewness	Kurtosis	[a]Lilliefors
Clay								
	Forward	4	0.324	0.409	0.2871	−0.252	3.201	0.9187
	Backward	7	0.319	0.406	0.2974	−0.223	3.276	0.8696
	Stepwise	4	0.324	0.409	0.2871	−0.252	3.201	0.9187
Sand								
	Forward	4	0.294	0.365	0.1984	0.080	2.946	0.7153
	Backward	6	0.293	0.358	0.2298	0.111	2.865	0.3559
	Stepwise	4	0.294	0.365	0.1984	0.080	2.946	0.7153
Silt								
	Forward	5	0.438	0.566	0.2343	−0.670	3.616	0.0833
	Backward	5	0.435	0.565	0.2363	−0.563	3.496	0.0618
	Stepwise	4	0.443	0.569	0.2246	−0.513	3.372	0.2074

[a]Lilliefors normality test (Kolmogorov-Smirnov)

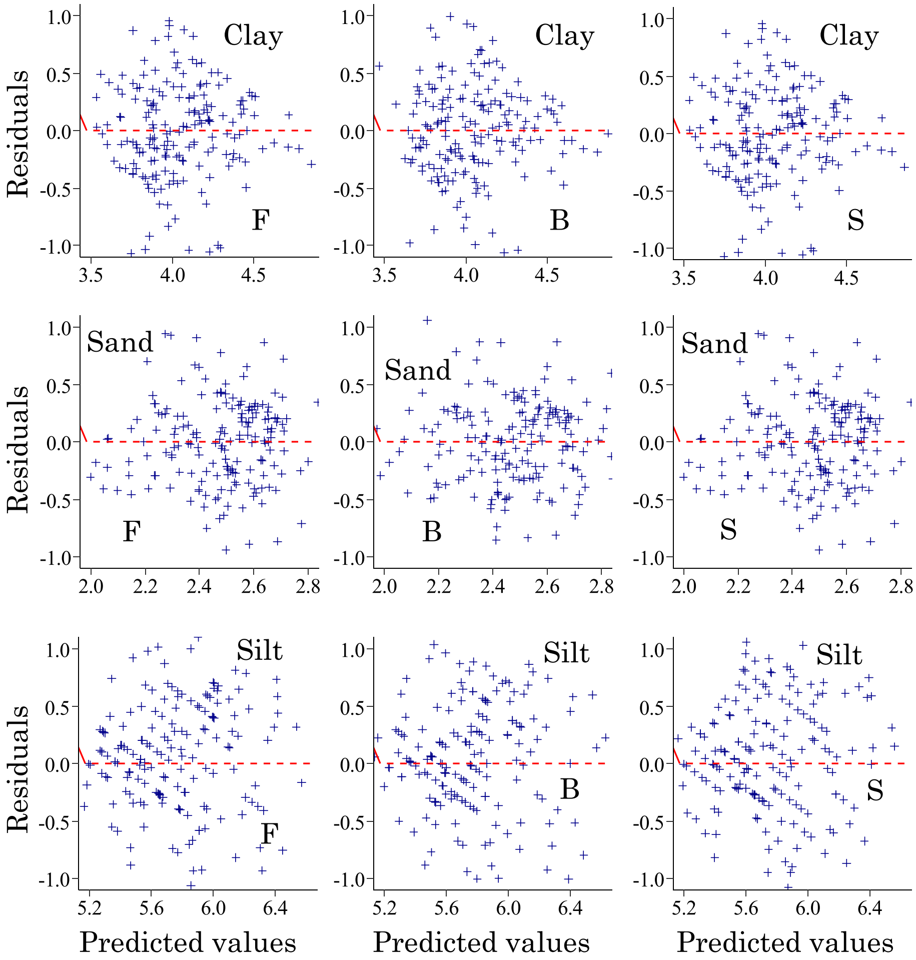

Fig. 2 Residuals plots of multiple linear regression models applying forward (F) selection, backward (B) elimination, and stepwise (S) selection methods

elimination method had the least MAE and RMSE for each of the soil texture fractions, and, consequently, it has the highest R^2 (Table 3). Figure 2 shows scatter plots of residuals and predicted values for homoscedasticity examination. As can be observed in any of the residuals plots, homoscedasticity was not violated by any of the regression models applying forward selection, backward elimination, and stepwise selection methods for clay, sand, and silt texture fractions as linearity (red lines) assumption was met having no obvious pattern wherein the residuals scattered randomly and symmetrically clustering towards the middle of the plot. For a graphic examination of the normality of residuals, Fig. 3 shows QQ-plots of residuals of the MLR models applying forward selection, backward elimination, and stepwise selection methods. Graphically, different selection methods revealed similar distribution of residuals where, generally noticeable, most of the points fall closely either below or above the reference line (red), and no obvious presence of outliers. The regression models for sand fraction have the most points that lie very close to the reference (red) line. This was also statistically confirmed in cross validation statistics (Table 3) where sand fraction regression models have the least skewness and kurtosis values. Nevertheless, none of the residuals distribution deviates greatly from normality and that residuals of all regression models are statistically normally distributed based on Lilliefors normality test (Table 3).

Relative importance of predictors

Figure 4 shows that stream proximity, elevation, mountain ridge proximity, channel network base level, flow accumulation, band 4, band 11, catchment area, vertical overland flow distance, overland flow distance, slope, and curvature classification were the significant predictors in this study.

Stream proximity was found to be the most important predictor in this study. Pahlavan-Rad and Akbarimoghaddam [8] found stream proximity or distance from the river as the most important variable in the spatial variability of soil texture fractions in a flood plain. For clay fraction, elevation, mountain ridge proximity, and

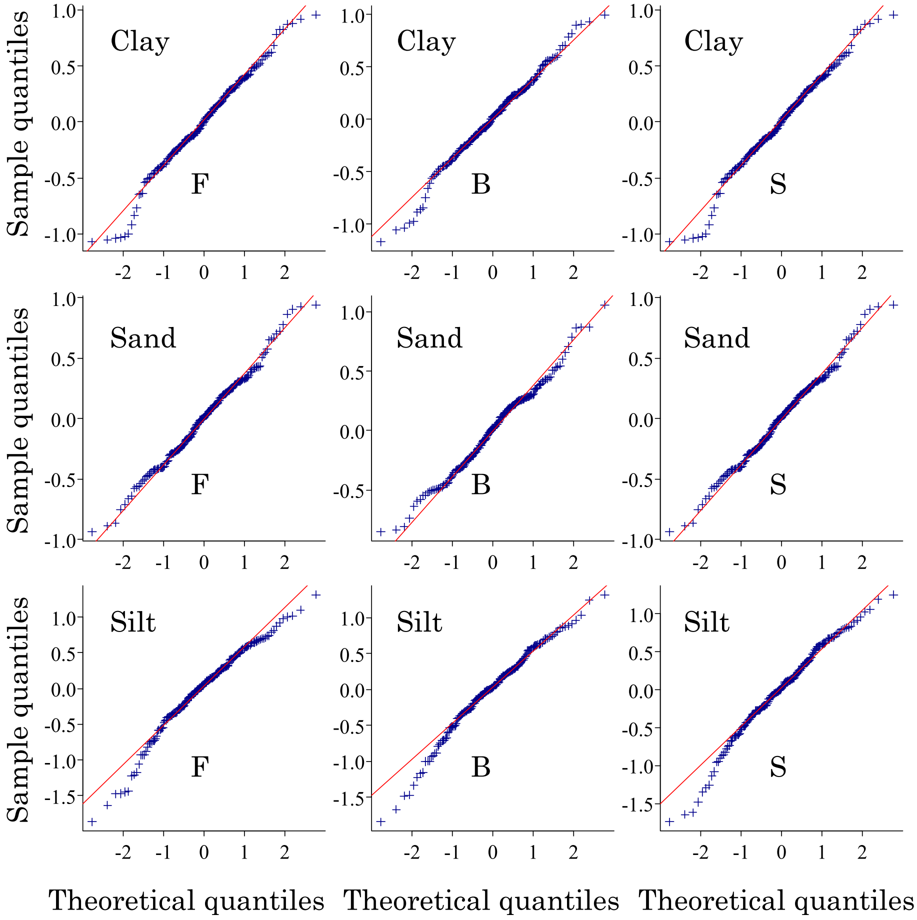

Fig. 3 QQ plots of residuals of the multiple linear regression models applying forward (F) selection, backward (B) elimination, and stepwise (S) selection methods

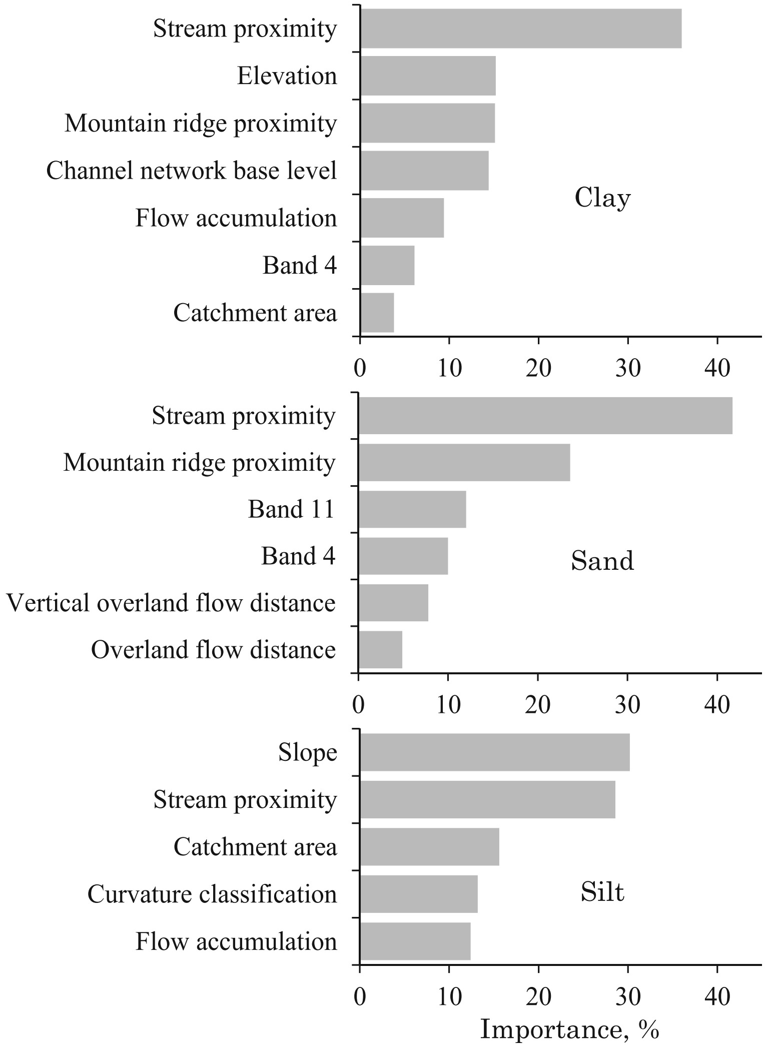

Fig. 4 Relative importance of predictors

channel network base level were about equally second most important predictors. Likewise, slope and stream proximity were about equally important for silt. Mountain ridge proximity or distance to the mountain ridge is a predictor probably first introduced by this study for DSM for soil texture fractions prediction. Mountain ridgeline is a peak elevation line that delineates watersheds at opposite sides of a mountain (leeward and wayward). The results suggest that stream proximity, mountain ridge proximity, and slope are the most important predictors that influence topsoil texture spatial variation, particularly for clay and sand fractions, in the study area. Remotely sensed data, such as Landsat 8, where band 4 and band 11 were found important, pointed out that optical imagery are potential auxiliary variables in enhancing prediction of topsoil texture fractions [37]. Flow accumulation, as in the study conducted by Castro-Franco et al. [22], was also the least important in the spatial variability of silt fraction at the southern Argentine Pampas.

Geospatial interpolation

After knowing significant predictors for each of the soil texture fractions (clay, sand, and silt), predictions of soil texture fractions were further evaluated applying kriging techniques. Likewise, linear regression was enhanced by combining regression with kriging. Prior to kriging, for every soil fraction, the experimental semivariogram was determined to examine the spatial dependence of dependent variables within the observed data points [6]. As a variogram model was required for geo-statistical interpolation, the variogram models that related to the best results in cross validation were the best fitted variogram models correspondingly for clay, sand, and silt. Hence, predefined models such as exponential, Gaussian,

Table 4 The variogram parameters of soil texture fractions applying exponential, Gaussian and exponential models

Soil Texture fraction	Semivariogram model	Nugget effect, C_0	Sill $[C_0 + C]$	Nugget-to-sill ratio $(C_0/[C_0 + C])$	Structured part-to-sill ratio $(C/[C_0 + C])$	Effective range
Clay						
	exponential	0.04	0.28	0.15	0.85	1560
	Gaussian	0.06	0.28	0.22	0.78	1900
	spherical	0.03	0.28	0.11	0.89	3800
Sand						
	exponential	0.03	0.21	0.14	0.86	2360
	Gaussian	0.05	0.21	0.24	0.76	2920
	spherical	0.03	0.21	0.14	0.86	5600
Silt						
	exponential	0.04	0.48	0.08	0.92	1250
	Gaussian	0.08	0.48	0.17	0.83	1570
	spherical	0.04	0.48	0.08	0.92	3490

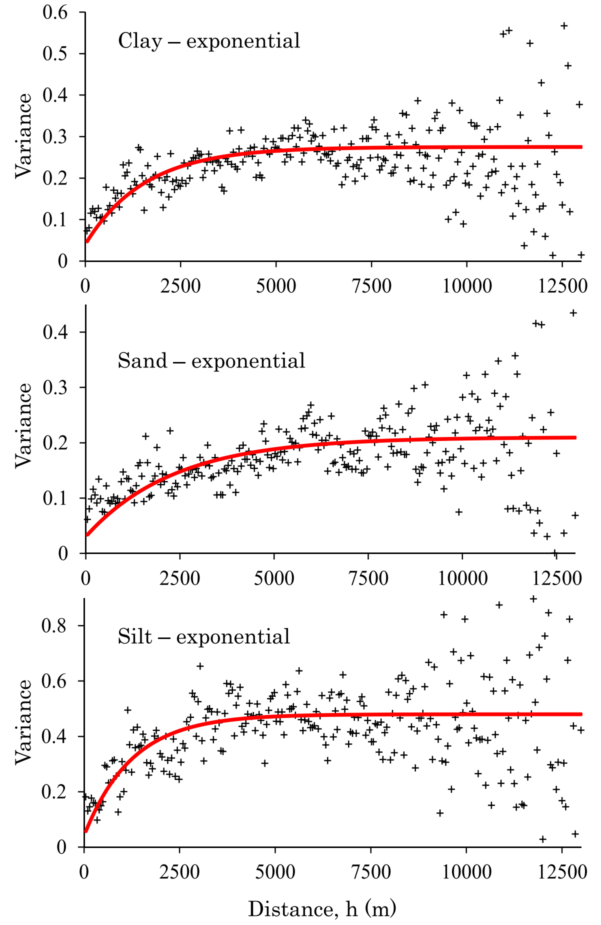

Fig. 5 Best fitted variogram models (red) for clay, sand, and silt

and spherical were fitted to the empirical semivariograms (Table 4 and Fig. 5). The mentioned variography was done in SAGA applying its variogram module.

Semivariogram

The nugget to sill (N:S, $C_0/[C_0 + C]$) ratio or its complementary structured part to sill ratio is considered as a spatial dependency criterion of a semivariogram [6]. The smaller the N:S ratio or the higher the structured part to sill ratio, the stronger the spatial dependency or autocorrelation [6, 55]. The N:S ratios for clay, sand, and silt varies within 0.11–0.22, 0.14–0.24, and 0.08–0.17, respectively, for the three variogram models (Table 4). These indicated that topsoil texture fractions have a strong spatial dependency, especially for silt, considering that N:S ratios are lesser than 0.25 [6, 55, 56]. Nugget effects of exponential and spherical models were lesser than the Gaussian model. Effective range of exponential, Gaussian, and spherical were ascending in order. Effective ranges of exponential and Gaussian models do not differ greatly, while effective range of spherical model is more or less twice the Gaussian model.

Performance of kriging techniques

The variogram parameters, as shown in Table 4, were adapted for OK, SK, and UK. All linear regression models considered in this study were coupled with UK. In the same manner, the performance of OK, SK, and UK applying exponential, Gausian, and spherical variaogram models were cross validated in terms of MAE, RMSE, and R^2 separately for clay, sand, and silt texture fractions (Tables 5, 6 and 7). Since forward and stepwise selection methods revealed the same significant predictors for clay and, likewise, for sand for the MLR (Table 3), UK-Stepwise in Tables 5 and 6 also represent UK-Forward for clay and sand, respectively.

Table 5 Cross validation of OK, SK, and UK applying exponential, Gaussian, and spherical variogram models applied to Clay texture fraction

Spatial Interpolation	Variogram model	MAE after kriging	RMSE	R^2
OK				
	exponential	0.130	0.186	0.8741
	Gaussian	0.234	0.312	0.5917
	spherical	0.134	0.191	0.8631
SK				
	exponential	0.130	0.186	0.8732
	Gaussian	0.234	0.312	0.5903
	spherical	0.134	0.191	0.8625
UK – Stepwise				
	exponential	0.128	0.182	0.8736
	Gaussian	0.223	0.296	0.6288
	spherical	0.133	0.187	0.8639
UK – Backward				
	exponential	0.126	0.178	0.8780
	Gaussian	0.219	0.289	0.6463
	spherical	0.131	0.183	0.8690

Table 6 Cross validation of OK, SK, and UK applying exponential, Gaussian, and spherical variogram models applied to Sand texture fraction

Spatial Interpolation	Variogram model	MAE	RMSE	R^2
OK				
	exponential	0.141	0.190	0.8198
	Gaussian	0.253	0.318	0.3982
	spherical	0.161	0.210	0.7741
SK				
	exponential	0.142	0.190	0.8199
	Gaussian	0.254	0.318	0.3980
	spherical	0.161	0.210	0.7742
UK – Stepwise				
	exponential	0.139	0.187	0.8160
	Gaussian	0.245	0.305	0.4429
	spherical	0.157	0.206	0.7718
UK – Backward				
	exponential	0.138	0.183	0.8213
	Gaussian	0.242	0.299	0.4640
	spherical	0.155	0.202	0.7788

Table 7 Cross validation of OK, SK, and UK applying exponential, Gaussian, and spherical variogram models applied to Silt texture fraction

Spatial Interpolation	Variogram model	MAE	RMSE	R^2
OK				
	exponential	0.144	0.230	0.8868
	Gaussian	0.279	0.375	0.6835
	spherical	0.178	0.259	0.8607
SK				
	exponential	0.144	0.230	0.8864
	Gaussian	0.278	0.375	0.6822
	spherical	0.177	0.259	0.8601
UK – Forward				
	exponential	0.141	0.221	0.8926
	Gaussian	0.268	0.345	0.7308
	spherical	0.172	0.247	0.8689
UK – Backward				
	exponential	0.141	0.221	0.8929
	Gaussian	0.267	0.345	0.7311
	spherical	0.172	0.246	0.8691
UK – Stepwise				
	exponential	0.141	0.221	0.8926
	Gaussian	0.268	0.344	0.7324
	spherical	0.172	0.246	0.8690

Generally, kriging techniques (Tables 5, 6 and 7) have better prediction accuracy than the linear regression models (Table 3) in terms of MAE, RMSE, and R^2. The performance of kriging techniques fitted with exponential or spherical variogram models was comparable, while kriging techniques fitted with Gaussian models were likewise at par (Tables 5, 6 and 7). As observed, kriging techniques fitted with Gaussian models have low prediction accuracy compared to kriging techniques fitted with exponential or spherical variogram models. Comparing MLR with kriging techniques, the respective MAE of clay_{MLR}, sand_{MLR}, and silt_{MLR} were about 2.5, 2, and 2.8 times as high compared to kriging techniques fitted with exponential or spherical variogram models. Similarly, MAEs of clay_{MLR}, sand_{MLR}, and silt_{MLR} were about 1.4, 1.2, and 1.6 times as high compared to kriging techniques fitted with Gaussian variogram model. In other words, R^2 values of kriging techniques with exponential or spherical variogram model were about 3, 3.8, and 3.8 times as high compared to clay_{MLR}, sand_{MLR}, and silt_{MLR}, respectively, while, R^2 results of kriging techniques with Gaussian variogram model were about 2.1, 2, and 3.1 times as high compared to clay_{MLR}, sand_{MLR}, and silt_{MLR}. Kriging technique with Gaussian variogram model had the least performance among the different kriging techniques applied in this study.

For clay texture fraction, UK with predictors by backward elimination method and exponential variogram model ($\text{Clay}_{\text{UK-back-expo}}$) revealed the lowest MAE and RMSE of 0.126 and 0.178, respectively. Consequently,

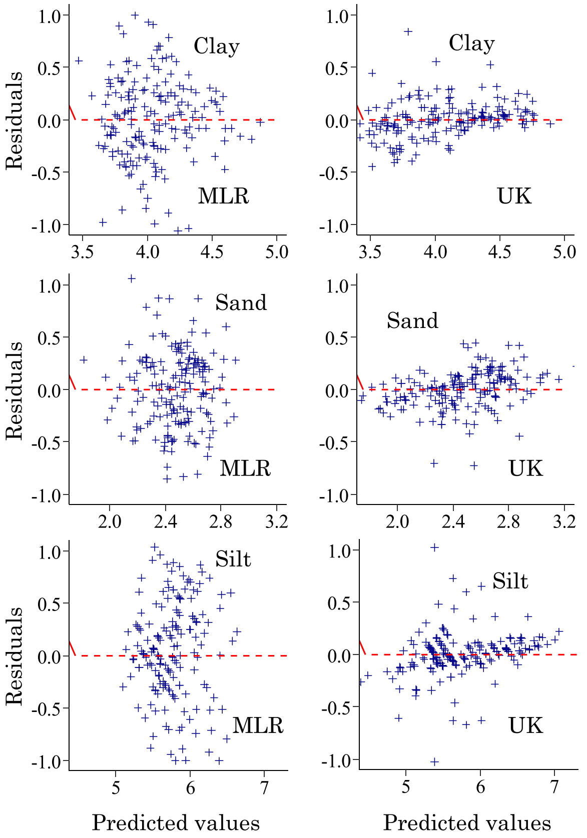

Fig. 6 Homoscedasticity and improvement of residuals from multiple linear regression (MLR) with backward selection method to universal kriging (UK) with exponential variogram applied to Clay, Sand, and Silt texture fractions

Clay-$_{\text{UK-back-expo}}$ exhibited the highest R^2 of 0.878 among the models presented in (Table 5). Moreover, graphical evaluation of residuals revealed significant improvement from MLR to UK. Residuals by UK reduced significantly closer towards the red line (Fig. 6). Thus, Clay$_{\text{UK-back-expo}}$ was considered the final model for transformed clay texture fraction as shown in Fig. 7, while Fig. 8 represents the final model of clay after back transformation.

Likewise, for sand texture fraction, UK with predictors by backward elimination method and exponential variogram model (Sand$_{\text{UK-back-expo}}$) revealed the lowest MAE and RMSE of 0.138 and 0.183, respectively. Consequently, Sand$_{\text{UK-back-expo}}$ exhibited the highest R^2 of 0.821 among the models presented in (Table 6). Furthermore, graphical visualization of residuals has shown significant improvement from MLR to UK. Residuals by UK reduced significantly closer towards the red line (Fig. 6). Thus, Sand$_{\text{UK-back-expo}}$ was considered the final model for sand texture fraction as shown in Fig. 9, while Fig. 10 represents the final model of sand after back transformation.

Similarly, for silt texture fraction, UK with predictors by backward elimination method and exponential variogram model (Silt$_{\text{UK-back-expo}}$) revealed the lowest MAE and RMSE of 0.141 and 0.221, respectively. Consequently, Silt$_{\text{UK-back-expo}}$ exhibited the highest R^2 of 0.8929 among the models presented in (Table 7). Moreover, graphical evaluation of residuals revealed significant improvement from MLR to UK. Residuals by UK reduced significantly closer towards the red line (Fig. 6). Thus, Silt$_{\text{UK-back-expo}}$ was considered the final model for silt texture fraction as shown in Fig. 11, while Fig. 12 represents the final model of silt after back transformation.

For the three soil texture fractions of clay, sand, and silt, UK with predictors by backward elimination method and exponential variogram model performed the highest predictive accuracy. This is an indication that the datasets are characterized by a linear trend, thus, UK is recommended to attain more fitting mapping results [44]. Nevertheless, performance of UK with predictors by stepwise selection method and exponential variogram model for clay, sand, and silt texture fractions was at par as shown from Tables 5, 6 and 7. Thus, UK can be considered a hybridized geospatial interpolation technique that coupled significant predictors from regression model and variography of regression residuals [57, 58]. However, in general, performance of OK and SK was comparable to UK (Tables 5, 6 and 7). Hence, for a straightforward and immediate estimate, OK and SK can be adapted [49]. Furthermore, it is observed that the stronger the spatial dependency or autocorrelation, the higher the R^2. Hence, high R^2 values of OK, SK, and UK were strongly attributed to high spatial dependency or autocorrelation.

Conclusions

Comprehending geographical distribution and accurate predictive mapping of topsoil texture fractions at a municipal scale are essential for watershed management, soil and water conservation, hydrological and crop suitability modeling. This fundamental soil texture geo-information processing approach is among the vital steps towards attaining a comprehensive land use plan. Spatial and geostatistics modeling, specifically MLR and kriging techniques (OK, SK, and UK), were fitted to predict topsoil texture fractions particularly for clay, sand, and silt. Data from sink-filled DEM derived environmental variables combined with remotely sensed images of Landsat 8 generated satisfactory results. Mountain ridge proximity, not considered among the predictors in DSM before but now, was found to be a significant predictor in this study. Further investigation should be conducted on the other side of the mountain ridge to understand better its predictive importance. Geospatial interpolations by kriging techniques were more accurate than MLR. The OK,

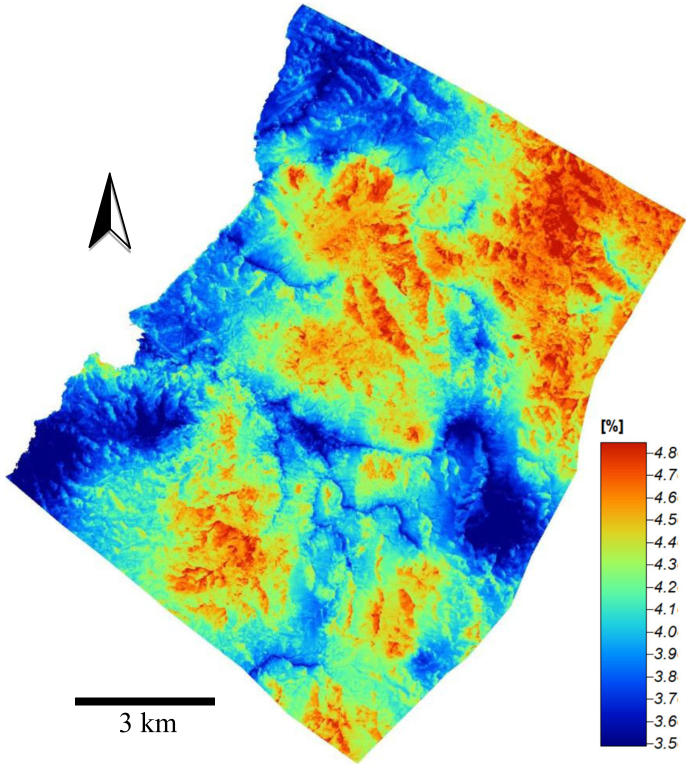

Fig. 7 Final Model of Transformed Clay by UK with predictors by backward elimination method and exponential variogram

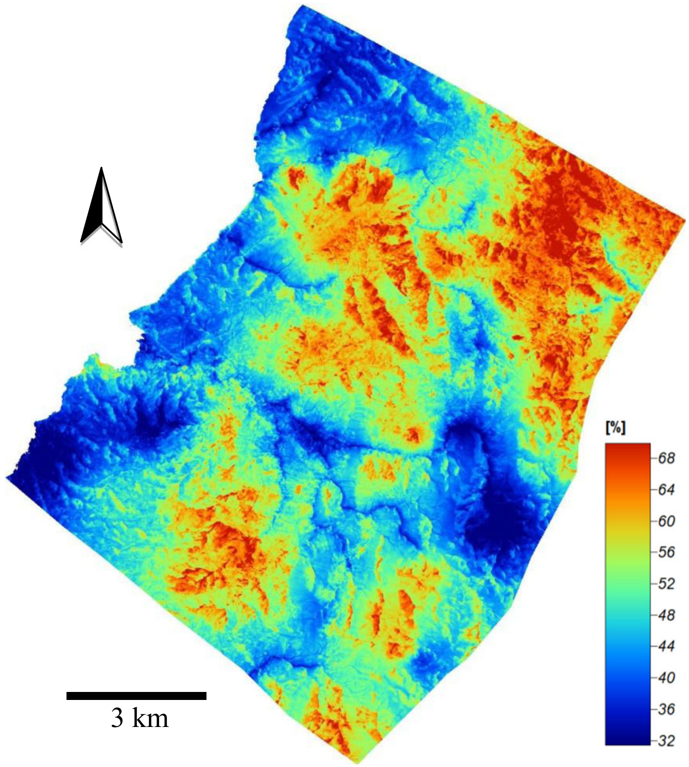

Fig. 8 Final Model of Clay after back transformation

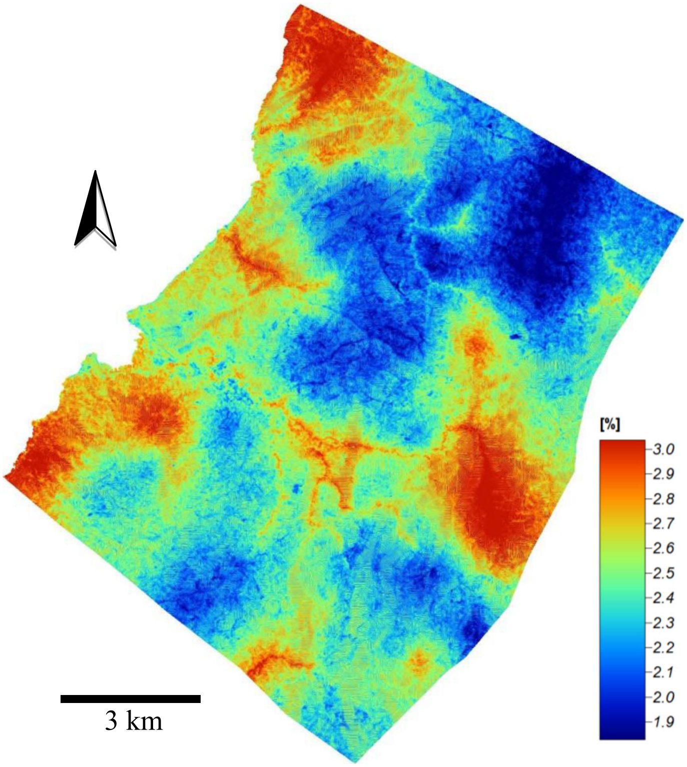

Fig. 9 Final Model of Transformed Sand by UK with predictors by backward elimination method and exponential variogram

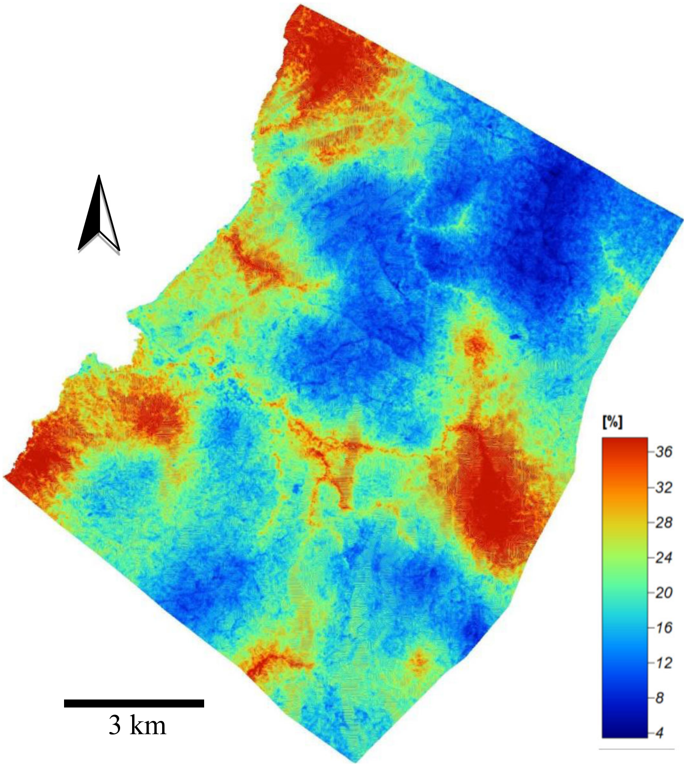

Fig. 10 Final Model of Sand after back transformation

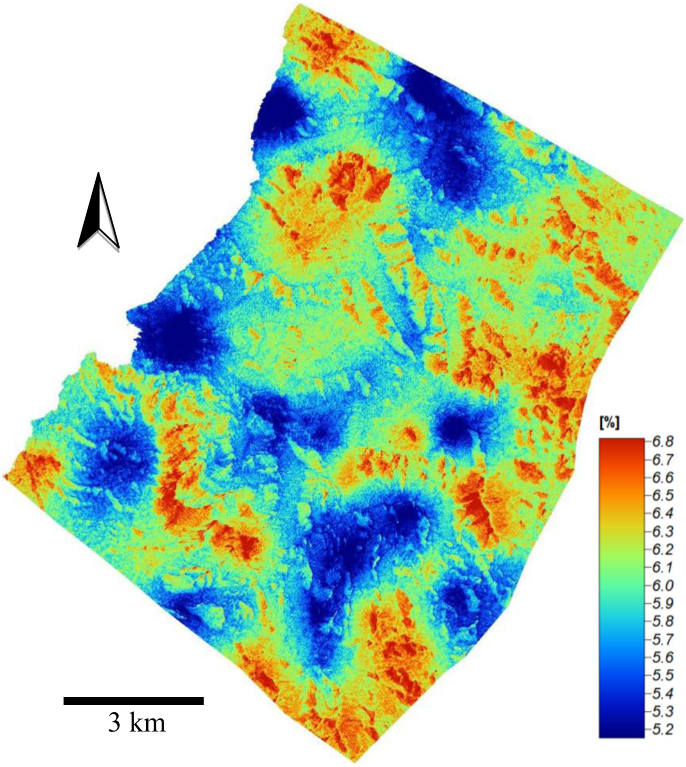

Fig. 11 Final Model of Transformed Silt by UK with predictors by backward elimination method and exponential variogram

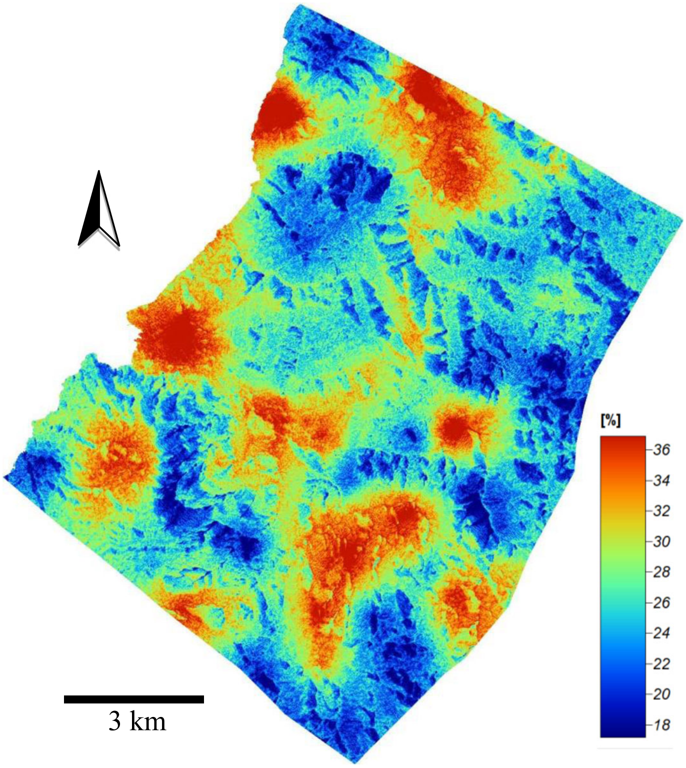

Fig. 12 Final model of Silt after back transformation

SK, and UK fitted with Gaussian variogram model were less suitable for predicting soil texture fractions whereas, UK coupled with predictors by backward elimination method and fitted with exponential variogram portrayed the most accurate in predicting topsoil texture fractions for clay, sand, and silt compared to any of the spatial and geostatistical modeling applied in this study. The results show that the transformation of dependent and independent variables provides homoscedasticity and normality of residuals to ensure unbiased estimation. With the significant cost advantage of this study on the use of FOSS, the models obtained by DSM revealed several benefits particularly in storing, retrieving and reproducing reliable geospatial information and ease of updating data and analysis. This study presents a successful approach in DSM with the use of SAGA GIS, QGIS, and R that are of best cost advantage to be adapted by any GIS users from a developing country like the Philippines. Moreover, additional soil samples and field validations of results are recommended for future research with the aim of improving accuracy and reliability of the models. Likewise, application of DSM should be extended in other municipalities throughout the central Visayas region and the rest of the country. The achieved methodology in this study is of great importance to decision makers at municipal, provincial, and regional scale for a very valuable outcome in achieving a more comprehensive land use plan since the generated results are useful for watershed management particularly for ecological, hydrological, and crop suitability modeling.

Acknowledgements

The authors are grateful to the U.S. Geological Survey Earth Resources Observation and Science Center (http://earthexplorer.usgs.gov/) for providing free images; to the UP Training Center for Applied Geodesy and Photogrammetry for providing free access of SAR DEM which is downloadable through https://lipad.dream.upd.edu.ph/. The authors also acknowledge the graduate scholarship funding from DOST-ERDT, Philippines that have made this research endeavour possible; to the University of San Carlos – Talamban Campus as the delivering institution and to the Cebu Technological University – Barili Campus as the sending institution.

Authors' contributions

JPM perceived and implemented the idea of applying digital soil mapping using open source software. JPM supervised soil sampling and facilitated soil analysis. AFT initiated image pre-processing. All authors performed the exploratory data analyses, data transformation, predictive modeling, and writing and revising of manuscript. All authors read and approved the final manuscript.

Funding

The Department of Science and Technology – Engineering Research and Development Technology (DOST-ERDT) provided the graduate scholarship fund that is course through the University of San Carlos – Talamban Campus as the delivering institution. The Cebu Technological University through its General Appropriations Act provided the funds for data collection and soil analysis. Interpretation of data, modeling, and writing of the manuscript were done by the authors.

Availability of data and materials

The datasets generated and analyzed in this study are submitted together with the manuscript. Such datasets are available from the corresponding author on reasonable request.

Competing interests

The authors declare they have no competing interests.

Received: 26 June 2019 Accepted: 18 October 2019
Published online: 04 December 2019

References

1. McBratney A, Field DJ, Koch A. The dimensions of soil security. Geoderma. 2014;213:203–13.
2. Zhang GL, Liu F, Song XD. Recent progress and future prospect of digital soil mapping: a review. J Integr Agr. 2017;16:2871–85.
3. Koch A, McBratney A, Adams M, Field D, Hill R, Crawford J, et al. Soil security: solving the global soil crisis. Glob Policy. 2013;4:434–41.
4. Pinheiro HSK, de Carvalho W, Chagas CD, dos Anjos LHC, Owens PR. Prediction of topsoil texture through regression trees and multiple linear regressions. Rev Bras Cienc Solo. 2018;42:1–21.
5. Khalil RZ, Khalid W, Akram M. Estimating of soil texture using landsat imagery: a case study of Thatta tehsil, Sindh. In: IEEE International Geoscience and Remote Sensing Symposium. Beijing; 2016 Jul 10–15.
6. Hosseini SZ, Kappas M, Bodaghabadi MB, Chahouki MAZ, Khojasteh ER. Comparison of different geostatistical methods for soil mapping using remote sensing and environmental variables in Poshtkouh rangelands. Iran Pol J Environ Stud. 2014;23:737–51.
7. Zhu AX, Hudson B, Burt J, Lubich K, Simonson D. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci Soc Am J. 2001;65:1463–72.
8. Pahlavan-Rad MR, Akbarimoghaddam A. Spatial variability of soil texture fractions and pH in a flood plain (case study from eastern Iran). Catena. 2018;160:275–81.
9. de Carvalho W, Lagacherie P, Chagas CD, Calderano B, Bhering SB. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. Geoderma. 2014;232–4:479–86.
10. Moller M, Volk M. Effective map scales for soil transport processes and related process domains - statistical and spatial characterization of their scale-specific inaccuracies. Geoderma. 2015;247–8:151–60.
11. Bhunia GS, Shit PK, Maiti R. Comparison of GIS-based interpolation methods for spatial distribution of soil organic carbon (SOC). J Saudi Soc Agric Sci. 2018;17:114–26.
12. Shit PK, Bhunia GS, Maiti R. Spatial analysis of soil properties using GIS based geostatistics models. Model Earth Syst Env. 2016;2:1–6.
13. Ließ M, Glaser B, Huwe B. Uncertainty in the spatial prediction of soil texture comparison of regression tree and random Forest models. Geoderma. 2012; 170:70–9.
14. Galvez JK. DA to produce new soils map in 45 days. The Manila Times; 2016.
15. Tejada SQ, Carating RB. Status of digital soil mapping in the Bureau of Soils and Water Management. In: Advancing the Science and Technology of Soil Information in Asia — Launch of the Global Soil Partnership's Asia Soil Science Network and GlobalSoilMap.net East Asia Node. Nanjing; 2012 Feb 8–11.
16. McDonald JH. Handbook of biological statistics. 3rd ed. Baltimore: Sparky House Publishing; 2014.
17. Mangiafico SS. Summary and analysis of extension program evaluation in R, version 1.18.1. New Brunswick: Rutgers Cooperative Extension; 2016.
18. Schmider E, Ziegler M, Danay E, Beyer L, Bühner M. Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. Methodology Eur. 2010;6:147–51.
19. Tsai AC, Liou M, Simak M, Cheng PE. On hyperbolic transformations to normality. Comput Stat Data An. 2017;115:250–66.
20. Zhang SW, Shen CY, Chen XY, Ye HC, Huang YF, Lai S. Spatial interpolation of soil texture using compositional kriging and regression kriging with consideration of the characteristics of compositional data and environment variables. J Integr Agr. 2013;12:1673–83.
21. Dempsey C. Computer specs for GIS work. Santa Clara: GIS Lounge; 2013.
22. Castro-Franco M, Domenech MB, Borda MR, Costa JL. A spatial dataset of topsoil texture for the southern argentine pampas. Geoderma Reg. 2018;12:18–27.
23. USGS. Landsat 8 (L8) data users handbook. Version 4.0. Reston: US Geological Survey; 2019.
24. Duarte L, Teodoro AC, Goncalves JA, Soares D, Cunha M. Assessing soil erosion risk using RUSLE through a GIS open source desktop and web application. Environ Monit Assess. 2016;188:351.

25. Dobarco MR, Orton TG, Arrouays D, Lemercier B, Paroissien JB, Walter C, et al. Prediction of soil texture using descriptive statistics and area-to-point kriging in region Centre (France). Geoderma Reg. 2016;7:279–92.
26. de Smith MJ, Goodchild MF, Longley PA. Geospatial analysis: a comprehensive guide to principles, techniques and software tools. 6th ed. Edinburgh: The Winchelsea Press; 2018.
27. Congedo L. Semi-automatic classification plugin documentation. Release 5. 0.0.1. 2016.
28. USGS. Landsat 7 (L7) data users handbook. Reston: US Geological Survey; 2018.
29. Mondejar JP, Tongco AF. Near infrared band of Landsat 8 as water index: a case study around Cordova and Lapu Lapu City, Cebu, Philippines. Sustain Environ Res. 2019;29:16.
30. Samira I, Ahmed D, Lhoussaine M. Soil fertility mapping: comparison of three spatial interpolation techniques. Int J Eng Res Technol. 2014;3:1635–43.
31. Webster R, Oliver MA. Geostatistics for environmental scientists. Chichester: John Wiley & Sons, Ltd; 2007.
32. Osborne JW. Best practices in data cleaning: a complete guide to everything you need to do before and after collecting your data. Los Angeles: SAGE Publications, Inc.; 2013.
33. Scott DW. Tukey ladder of powers. In: lane DM, editor. Introduction to statistics. Online Statistics Education: A Multimedia Course of Study. http://onlinestatbook.com/Online_Statistics_Education.pdf. Accessed 21 Dec 2018.
34. Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, et al. System for automated Geoscientific analyses (SAGA) v. 2.1.4. Geosci Model Dev. 2015;8: 1991–2007.
35. Gromping U. Relative importance for linear regression in R: the package relaimpo. J Stat Softw. 2006;17:1–27.
36. Lindeman RH, Merenda PF, Gold RZ. Introduction to bivariate and multivariate analysis. Glenview: Scott Foresman & Co; 1980.
37. Liao KH, Xu SH, Wu JC, Zhu Q. Spatial estimation of surface soil texture using remote sensing data. Soil Sci Plant Nutr. 2013;59:488–500.
38. Wang Z, Shi WJ. Mapping soil particle-size fractions: a comparison of compositional kriging and log-ratio kriging. J Hydrol. 2017;546:526–41.
39. Sari RKN, Pasaribu US. The comparison of isotropic and anisotropic semivariogram for gauss model. AIP Conf Proc. 2014;1589:508–11.
40. Hengl T. A practical guide to geostatistical mapping of environmental variables. Luxembourg: European Communities; 2007.
41. Delbari M, Afrasiab P, Loiskandl W. Geostatistical analysis of soil texture fractions on the field scale. Soil Water Res. 2011;6:173–89.
42. de Menezes MD, Silva SHG, de Mello CR, Owens PR, Curi N. Spatial prediction of soil properties in two contrasting physiographic regions in Brazil. Sci Agr. 2016;73:274–85.
43. Shaffer JM. The effects of spatial resolution on digital soil attribute mapping [Master's thesis]. Columbus: The Ohio State Univ; 2013.
44. Kiš IM. Comparison of ordinary and universal Kriging interpolation techniques on a depth variable (a case of linear spatial trend), case study of the Šandrovac Field. Min Geol Petrol Eng Bull. 2016;31:41–58.
45. Haldar SK. Statistical and geostatistical applications in geology. In: Haldar SK, editor. In: mineral exploration: principles and applications. Waltham: Elsevier; 2013. p. 157–82.
46. Daya AA, Bejari H. A comparative study between simple kriging and ordinary kriging for estimating and modeling the cu concentration in Chehlkureh deposit. SE Iran Arab J Geosci. 2015;8:6003–20.
47. Mpanza M. A comparison of ordinary and simple Kriging on a PGE resource in the eastern limb of the Bushveld complex [Master's thesis]. Johannesburg: Univ of the Witwatersrand; 2015.
48. Lichtenstern A. Kriging methods in spatial statistics [Bachelor's thesis]. Munich: Technical Univ of Munich; 2013.
49. Hengl T, Heuvelink GBM, Stein A. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma. 2004;120:75–93.
50. Ballabio C, Panagos P, Monatanarella L. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma. 2016;261:110–23.
51. Kuhn M, Johnson K. Applied predictive modeling. 1st ed. New York: Springer; 2013.
52. Yang RM, Liu F, Zhang GL, Zhao YG, Li DC, Yang JL, et al. Mapping soil texture based on field soil moisture observations at a high temporal resolution in an oasis agricultural area. Pedosphere. 2016;26:699–708.
53. Zhang LM, Liu YL, Li XD, Huang LB, Yu DS, Shi XZ, et al. Effects of soil map scales on simulating soil organic carbon changes of upland soils in eastern China. Geoderma. 2018;312:159–69.
54. Heil K, Schmidhalter U. Improved evaluation of field experiments by accounting for inherent soil variability. Eur J Agron. 2017;89:1–15.
55. Omran ELE. Improving the prediction accuracy of soil mapping through geostatistics. Int J Geosci. 2012;3:574–90.
56. Hu XS, Xu HQ. A new remote sensing index for assessing the spatial heterogeneity in urban ecological quality: a case from Fuzhou City. China Ecol Indic. 2018;89:11–21.
57. Ma YX, Minasny B, Wu CF. Mapping key soil properties to support agricultural production in eastern China. Geoderma Reg. 2017;10:144–53.
58. Wackernagel H. Multivariate geostatistics: an introduction with applications. 3rd ed. Berlin: Springer; 2003.

Publisher's Note

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