Soil Organic Matter Prediction with Sentinel-2 and Geostatistical Methods

Tessnika Sewpersad, Sifiso Xulu* and Michael Gebreslasie
School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Westville Campus, Durban 4000, South Africa
*Corresponding author, email: xulus4@ukzn.ac.za, Tel (+27) 031 260 7656

DOI: https://dx.doi.org/10.4314/sajg.v14i2.3

Abstract

Predicting soil organic matter (SOM) using advanced tools such as Sentinel-2 satellite imagery and geostatistical methods could help in achieving the Sustainable Development Goals related to food security and climate change. The spectral capabilities of Sentinel-2 allow for the characterization of soil properties from the reflectance patterns of soil, thereby allowing for indirect SOM estimation. This is imperative for countries, such as South Africa, which have low SOM levels and therefore limited SOM data. Okhahlamba Local Municipality in uThukela District Municipality, KwaZulu-Natal Province of South Africa, was chosen as a testing site in this study. Here, we used stratified random sampling; 52 samples were selected − 13 in each of the main land-use classes (agricultural, residential, rangeland and eroded) in the municipal area. Sentinel-2 data were used in conjunction with the geostatistical methods, Ordinary Kriging (OK) and Simple Kriging (SK), and the hybrid geostatistical methods, Regression Ordinary Kriging (ROK) and Regression Simple Kriging (RSK), to predict SOM. The performance of the hybrid methods proved to be superior to that of the ordinary methods. Specifically, ROK and RSK, using spectral bands, presented with the highest levels of accuracy (R^2: 0.63 for both) and the lowest proportion of errors (ME: 0.53%; 0.54%) and (RMSE: 0.68%; 0.67%), respectively. The ROK andR SK outputs emanating from further methods using principal components (R^2: 0.30; 0.26), (ME: 0.70%; 0.71%) and (RMSE: 0.82%;0.85%). OK and SK showed the lowest levels of accuracy (R^2: 0.07; 0.04) and the largest proportion of errors (ME: 0.79%; 0.83%) and (RMSE: 0.93%; 0.98%). Auxiliary information served to boost the predictive performance of these models. Overall, ROK and RSK, using spectral bands, accounted for 63% of the predicted SOM variability. SOM plays a crucial role in soils and should be sustainably managed in order to improve food security for the current and future generations. The combined use of Sentinel-2 data and geostatistical methods has great potential for predicting SOM and thereby in achieving the sustainable development agenda.

Keywords: *Soil organic matter, regression kriging, prediction, geostatistics*

1. Introduction

Soil organic matter (SOM), which is composed of soil biota and biological residues at all levels of decay (Craswell and Lefroy, 2001), is the key indicator of soil health and has major impacts on land-based resources. According to Juma (1999), such decay produces organic carbon and nutrients, crucial components of SOM: in fact, they become a source of energy for soil microorganisms in that they break down organic material into forms that plants can easily absorb. From an agricultural perspective, SOM serves as a complex reservoir of carbon, nutrients, and microorganisms that promotes soil fertility and improves overall soil health and function (Amulothu *et al.*, 2023). These benefits improve food production and security (Gattinger *et al.*, 2011) and, as reflected in a growing number of studies (Lorenz and Lal, 2018; Smith *et al.*, 2018; Lal, 2020a), ensure that SOM is an essential component in promoting the United Nations' Sustainable Development Goals (SDGs). However, such advantages require a series of stra tegies to improve the understanding of SOM and of the robust and consistent methods to qualify it.

Beyond the appropriate effects of SOM on agricultural activities, its importance also includes its role in mitigating climate change. This fact was strengthened at the COP21 conference in 2015, where the "Four per 1000 Soils for Food Security and Climate" programme was introduced. This programme recognizes that soils have a critical role to play in mitigating greenhouse gas (GHG) emissions caused by human activity, in ensuring food security (Lal, 2020a). While the intention of this initiative is contentious (King *et al.*, 2018; Aubert *et al.*, 2020), numerous efforts related to the elements linked to it are ongoing in several parts of the world (Smith *et al.*, 2018). To promote the “Four per 1000 Soils” initiative, the evaluation and estimation of SOM is key to understanding the effects of land and soil management (Lal, 2020b). This is a challenging task because SOM content varies greatly from local to global landscapes (Pouladi *et al.*, 2023), and innovative digital soil mapping methods have revolutionized the spatial quantification and prediction of SOM (McBratney, Santos and Minasny, 2003).

SOM is likely to result in changes to the biophysiological properties of soil that affect their spectral reflectance (Li *et al.*, 2024), thereby making remote sensing a suitable tool to objectively quantify SOM conditions (Gallo *et al.*, 2018a). In particular, the several bands and their combinations in the visible (VIS), near-infrared (NIR), and shortwave infrared (SWIR) spectral bands are closely related to the organic and inorganic composition of the soil (Ben-Dor, Irons and Epema, 1999; Krishan *et al.*, 2014) and provide a theoretical basis for SOM investigations (Shepherd and Walsh, 2007). For example, Viscarra Rossel *et al.* (2006) showed good correlations of the SOM along the respective wavelengths, 410 nm, 570 nm, and 660 nm of the VIS region. Liu, Zhang and Zhang (2009) found that reflectance in the range of 620 nm to 810 nm is related to the SOM and that the highest correlation coefficient is at 710 nm. Also, Zheng *et al.* (2016) showed that the wavelength range, 500–700 nm, presented with the highest correlation with SOM, while Meng *et al.* (2020) found that the wavelength 700–800 nm is most useful for SOM prediction. The retrieval of SOM from these regions has played a major role in characterizing the spatial and temporal status of soil information dynamics, ranging from the local to larger areas (Gallo *et al.*, 2018b). This has revolutionized the long-standing challenge of SOM quantification using spatially- and cost-constrained field campaigns (Wang *et al.*, 2017), as newer remote sensing products such as Sentinel-2 offer better opportunities to estimate SOM using rapid, cost-free, spectrally rich and spatially continuous data (Zhang *et al.*, 2021). These features enable the detection of fine-scale SOM variations which is crucial for SOM prediction in various landscapes (Li *et al.*, 2021).

Numerous studies have integrated various remotely sensed products with different spatial modelling strategies to improve the reliability of SOM prediction. For example, Chang *et al.* (2024) successfully predicted SOM in croplands when using machine-learning algorithms. Wang *et al.* (2024) combined PlanetScope and Sentinel-2 for SOM prediction and achieved satisfactory results. Sentinel-2 data have appeared as the most popular product in SOM research recently in that, with the integration of geostatistical techniques, they employ a definite framework of spatial autocorrelation, and in so doing, produce an estimate of a variable at each point (McBratney, Santos and Minasny, 2003). Conversely, deterministic systems (e.g., splines) tend to generalize reality in that they fail to account for the spatial autocorrelation of samples, including the estimation of errors (Robinson and Metternicht, 2006). Therefore, lately, geostatistical methods, particularly the hybrid varieties, have become a preferred choice owing to their success in SOM prediction (Piccini, Marchetti and Francaviglia, 2014). For example, Yao *et al.* (2013) showed that regression kriging (RK) outperformed ordinary kriging (OK) and inverse distance weighting (IDW) in their study of SOM in a small-scale heterogeneous terrain. Bhunia, Shit, and Maiti (2018) successfully compared different deterministic and geostatistical methods for mapping soil organic content and found OK to be the best method. Bangroo *et al.* (2020) predicted soil organic content using RK and OK and achieved greater accuracy in their results when they used RK. Because of their theoretical applicability in this study and their superior performance in prior works, the geostatistical techniques, OK, SK, and the hybrid geostatistical methods, ROK and RSK were taken into consideration.

According to du Preez, van Huyssteen and Mnkeni (2011), South Africa is typically poor in organic soils, with much of them − owing to their relatively wide extent − concentrated in the country’s grasslands and savanna biomes, while on the other hand, the relatively small forest biome contains more organic soils per square metre (Schütte, Schulze and Paterson, 2019). Geographically, organic soils occur mainly at higher altitudes in the central-southern parts of KwaZulu-Natal, in the north-eastern part of the Eastern Cape, and in the southern parts of the Western Cape Province, with isolated patches in another places (Schütte, Schulze and Paterson, 2019). In this study, we predicted the proportion of SOM in the soils of the central western part of KwaZulu-Natal, South Africa, by using the geostatistical methods, Ordinary Kriging (OK) and Simple Kriging (SK), and the hybrid geostatistical methods, Regression Ordinary Kriging (ROK) and Regression Simple Kriging (RSK), based on Sentinel-2 data.

2. Materials and Methods

2.1. Study area

The Emakhosaneni area (latitude 28° 46' 45.1"S; longitude 29° 14' 42.4"E), located 17 km southwest of Bergville in the central section of the Okhahlamba Local Municipality in Kwa-Zulu-Natal province, was the site selected for the SOM prediction (Fig. 1). At the base of the Drakensberg range, the region is situated at an altitude of 1,150 metres and shares its boundaries with the province of the Free State and the Kingdom of Lesotho (Sewpersad *et al.*, 2024). With 650–1200 mm of annual precipitation, the majority of which falls between October and March, this area has a significantly seasonal rainfall pattern and is extremely humid (Schulze, 1997). Typically, the lowest temperature is 12°C, while the highest is 28°C (Schulze, 1997).

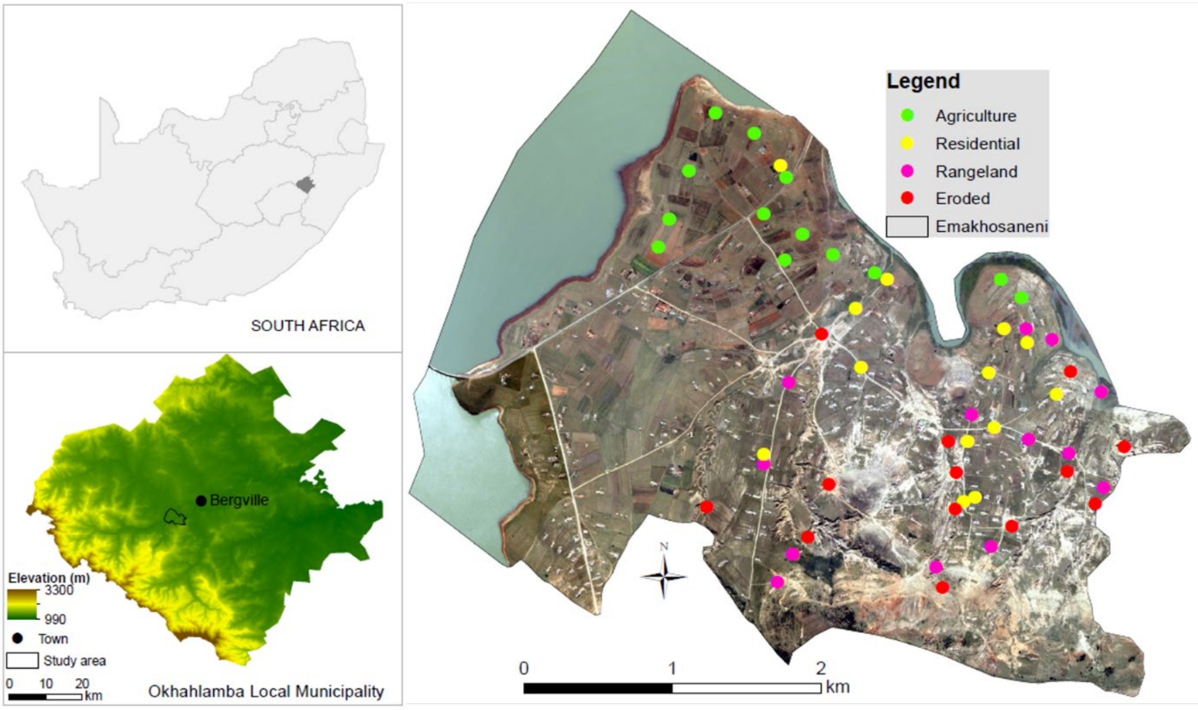

Fig. 1. Geographic location of the research site and the soil samples collected across the mainland-use categories.

The region is swept by strong winds, predominantly in the winter and spring periods. Numerous water sources are found in the area, with the main ones being the Tugela River and the Driel Dam to the east, and the Woodstock Dam to the north. The primary geology of the region is composed of Beaufort Group rock types that are Triassic and Permian in age (Verster, 1998). Occupying 13.72km^2, the region accommodated 1,938 residents overall, and had an average population density of 141.30 persons per km^2 in 2011 (Statistics South Africa (SSA),

2011) − which is now anticipated to be somewhat higher. More than 90% of the population is black African, and under traditional leadership, with the people preserving a vibrant Zulu culture. This region is devoted mainly to large- and small-scale agriculture, with livestock grazing the grasslands.

2.2. Soil sample collection

We collected soil samples on 16 July 2019, on the same date that the satellite imagery employed in this work was acquired. After a field campaign, we identified four main classes of land in the area: agricultural, residential, rangeland, and eroded. We then applied a stratified random sampling method, where 13 samples from each class were selected individually to make up 52 samples, the distribution of which is shown in Fig. 1. Using an auger, we collected samples from the surface to a depth of 15 cm[1] and in cases where this tool could not penetrate, we used a spade. We then placed the respective samples in labelled, airtight plastic bags to prevent any spillage or drying out and stored them in a cool,dry cooler box, away from direct exposure to sunlight. All 52 samples were stored in the laboratory for analysis. The GPS points for the samples were recorded in MS Excel and a complete data file was exported in *.csv* format for descriptive and geostatistical analysis.

2.3. Sentinel-2 spectral data

We used cloud-free Sentinel-2 satellite imagery taken on 16 July 2019. The preference was given to Sentinel-2 over other possible satellite products such as PlanetScope and SPOT. This was due to its rich spectral properties (13 bands), which are commonly used in SOM research (Castaldi *et al*., 2019). Sentinel-2 satellite imagery collects multispectral data at 10 to 60 m spatial resolutions across these bands (Drusch *et al.*, 2012). These data were obtained at Level 2A (in the lower layers of the atmosphere), with geometric, radiometric, and atmospheric correction, which together make the data suitable for analysis (Žížala, Minařík and Zádorová, 2019). The product was further processed by means of the Sentinel Application Platform (SNAP) programme (ESA, 2020)[2]. Subsequently, all selected bands were resampled to 10 m using a nearest neighbour algorithm to ensure that all bands had a similar resolution scale. This method is computationally efficient and is known to preserve the input pixel values of the image (Roy *et al.*, 2016; Gholizadeh *et al.*, 2018). We used nine of the thirteen bands, namely B2 (Blue), B3 (Green), B4 (Red), B5 (Red edge1), B6 (Red edge2), B7 (Red edge3), B8 (NIR), B11 (SWIR1) and B12 (SWIR2), all of which are commonly used to characterise soils (Gholizadeh *et al.*, 2018; Castaldi *et al.*, 2019).

---

[1] This is the commonly used depth of the soil in SOM studies because it represents the top layer, which is often the layer compared with the remotely sensed data in question.

[2] The Sentinel-2 user manual (ESA, 2015) provides an overview of the process.

2.4. Principal component analysis (PCA)

We used the PCA method, a widely applied multivariate statistical method in SOM research (Guo *et al.*, 2018) because it has a strong impact on predicting soil characteristics (Hengl, Heuvelink and Rossiter, 2007). In general, PCA minimizes the dimensionality of the dataset and avoids multicollinearity without losing any information from the original data. As such, the resultant PCs explain most of the information from the dataset (Sharma, Chauhan and Kumar, 2021). We implemented PCA by applying factor analysis in the Statistical Package for the Social Sciences (SPSS; version 26), where the computed spectral indices served as the input variables. Kaiser's criterion was applied to identify the dominant PCs with eigenvalues over one (1), which were then rotated, using varimax rotation, to maximize the correlations between the PCs and the variables (Askari *et al.*, 2020). Owing to the moderate nature of the multicollinearity and the fact that it fell within the brink (i.e., <10), we followed Samuels (2017) and implemented PCA only for selected spectral indices. Although PCA has been successfully used in soil quality research, there have been cases where it has fallen short in selecting indicators that would reflect the differences between soil conditions in different land systems (Nehrani *et al*., 2020).

2.5. Multiple linear regression (MLR)

We developed MLR models to predict SOM content using multivariate analysis techniques. According to Liu *et al.* (2015), MLR assumes that the independent and dependent variables are linearly related. Additionally, this method aims to quantify the total influence of independent factors on a particular dependent variable (Kumar *et al.*, 2018). Here, the response variable was SOM, with predictor variables consisting of PCs obtained from the spectral bands. Two distinct MLR models were constructed - one with nine spectral bands as predictor variables and SOM as the dependent variable. Significant descriptive statistics and residuals (differences between expected and observed SOM values) were analysed. In the second model, the PCA scores served as the predictor variables and SOM as the response variable. Predicted values and residuals from both models were stored for further examination.

MRL is computed in Equation 1.

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \dots\dots\dots\dots\dots\dots + b_pX_p \qquad [1]$$

Here, X denotes the dependent variable; $X_1 + X_2 + X_3 \dots\dots\dots X_p$ denote the independent variables; and b_0 represents the value of Y when all of the independent variables ($X_1 - X_p$) are equal to zero (i.e., the constant), and b_1 through b_p are the predicted regression coefficients. The individual regression coefficient represents the variation in Y relative to a one-unit change in the corresponding independent variable.

2.6. Geostatistical analysis

In this study, we used different geostatistical methods to predict SOM content in an ArcGIS (version 10.2) environment. This included Ordinary Kriging (OK) and Simple Kriging (SK) as commonly applied methods for this purpose. OK is capable of proving unbiased predictions with the least errors (Gia Pham *et al.*, 2019), while SK is relatively similar to OK except that it integrates a mean value in the prediction of values at unknown locations (Webster and Oliver, 2007).

OK (Equation 2) and SK (Equation 3) are calculated as:

$$Z(X_0) = \sum_{i=1}^{N} \lambda_i Z(X_i) \quad [2]$$

$$Z(X_0) = \sum_{i=1}^{N} \lambda_i \left(Z(X_i) - m(\mu_i)\right) \quad [3]$$

Here, $Z(X_0)$ signifies the predicted value at the unmeasured site X_0; $Z(X_i)$ represents the observed value at site (X_i); λi is the weighting coefficient from the observed site to X_0; $m(\mu_i)$ is the recognized stationary mean of $Z(X_i)$; and N represents the number of sites within the nearest neighbourhood searching process.

Beyond the above-mentioned methods, an attempt was made to incorporate a regression model with residuals to produce variants such as regression kriging (RK). Thus, RK integrates the strength of both the regression and kriging analyses to elevate the accuracy of geospatial predictions (Hengl, Heuvelink and Stein, 2004). Essentially, this is valuable for neighbouring datasets with comparable values, a process known as spatial autocorrelation (Suleymanov *et al.*, 2024). According to Kumar *et al.* (2018), RK entails the following steps: (a) Use an MLR to identify a trend between the response variable (SOM) and auxiliary variables (predictors; spectral bands, and PCs); and (b) Use regression techniques to reach a local mean and to estimate the associated residuals using a variogram and kriging, and subsequently add them all together. RK divides the prediction into two portions, as shown in Equation 4. The equation on the right-hand side is first represented by regression and the second part by kriging the residual.

$$Z(X_0) = \hat{m}(X_0) + \hat{e}(X_0) = \sum_{k=0}^{p} \widehat{\beta_k} \cdot q_k(X_0) + \sum_{i=1}^{N} \lambda_i \cdot e(X_i) \quad [4]$$

Here, $\hat{m}(X_0)$ represents the fitted deterministic portion; $\hat{e}(X_0)$ is the interpolated residual; $\widehat{\beta_k}$ is the predicted deterministic model coefficient; $q_k(X_0)$ is the kth predictor at site X_0; p is the number of predictors; λ_i is the kriging weight resulting from the geographic dependence pattern of the residual; and $e(X_i)$ is the residual at site X_i.

In this study, we performed both OK and SK methods, on the residuals obtained from the MLR model of the spectral bands by using the training dataset. After the residuals were interpolated, the projected values of the MLR models were added to them. Regression Ordinary Kriging (ROK) and Regression Simple Kriging (RSK) were the outcomes of this. Likewise, for the spectral indices model, the afore-mentioned process was carried out.

$$\gamma(h) = \begin{Bmatrix} 0 & h = 0 \\ C_0 + C\left(\frac{3h}{2h} - \frac{h^3}{a^3}\right) & 0 < h \leq a \\ C_0 & h > a \end{Bmatrix} \quad [5]$$

Here, $\gamma(h)$ represents the variogram; a represents the range of soil samples; h represents the spatial lag; C_0 is the nugget; and C_0+C is the partial sill. We used a semivariogram for the prediction process.

2.7. Validation

Before building the prediction models, we first split the dataset into training data (80%) and validation data (20%). According to Stevens and Ramirez-Lopez (2014), using independent training and validation samples makes it easier to compare SOM results between different models. We also tested the accuracy of the prediction methods (OK, SK, ROK, and RSK) for SOM prediction. We then validated the accuracy of each model by using the determination coefficient (R^2; Eq. 6), the mean error (ME; Eq. 7), and the root mean square error (RMSE; Eq. 8). In general, the model with the greatest R^2 and the least ME and RMSE values proved to be the most accurate.

Equations (6)–(8) are calculated as:

$$R^2 = \left(\frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}\right)^2 \quad [6]$$

$$ME = \frac{1}{N}\sum_{i-1}^{N}(z(x_i) - z^*(x_i)) \quad [7]$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i-1}^{N}(z(x_i) - z^*(x_i))^2} \quad [8]$$

Here, x_i represents the predicted SOM at site i; y_i denotes the measured SOM at site i; n represents the number of samples; $z * (x_i)$ is the predicted value; and $z(x_i)^2$ is the measured value.

3. Results

3.1. Evaluating SOM predictions with geostatistical techniques

A summary of the variables used to simulate the residuals of the two MLR models and the semivariograms for the SOM content are outlined in Table 1. As opposed to the residuals of the two distinct MLR models, the measured SOM content presented with greater values over the range of the semivariograms. More precisely, the lowest range (648 m) was achieved by ROK by using the residuals of the spectral band, while the maximum range (868.45 m) was covered by the semivariogram for SK which used the SOM content. The nugget ranged from 0.390 to 1.013, partial sill from 0.365 to 0.736, and sill 0.885 to 1.446. Furthermore, the nugget-to-sill ratio, within the remaining ratios in the region between 0.458 and 0.511, was greatest for the SOM content evaluated via OK (0.735) and smaller for SK (0.406).

Table 1. Semivariogram model variables for residuals of the MLR models obtained from PCs and spectral bands, individually, and SOM content

Data type	Method	Range (m)	Nugget (C_0)	Partial sill (C)	Sill ($C0 + C$)	C_0/Sill
Measured SOM	OK	723.21	1.013	0.365	1.378	0.735
Measured SOM	SK	868.45	0.390	0.570	0.960	0.406
Residuals of spectral bands[1]	ROK	648.00	0.452	0.433	0.885	0.511
Residuals of spectral bands[1]	RSK	785.29	0.570	0.604	1.174	0.486
Residuals of PCs[1]	ROK	690.87	0.710	0.736	1.446	0.491
Residuals of PCs[1]	RSK	814.57	0.511	0.604	1.115	0.458

[1]Obtained from the MLR model

The scatterplots comparing the predicted SOM values from the different geostatistical models with the observed SOM for the training and validation datasets are shown in Fig. 2. Figs 2(a) and (b) show a roughly continuous trend because the lines of best fit have a horizontal pattern. Except for a few cases, the plotted points are dispersed randomly in both scenarios and are very near to their corresponding best-fit lines. The plots of the training and validation datasets, however, clearly show a linear trend in Figs 2(c) and (d) since the best-fit lines are virtually at a 45° angle, with just a small number of points deviating from it. As a result, it is clear that the plotted points are closer to their individual lines of best fit and thus more compact. Additionally, a linear trend was shown for the training and validation datasets in Figs 2(e) and (f). But more plots were straying from the optimal fit lines. As a result, the data points were more dispersed along their respective lines, and therefore less compact.

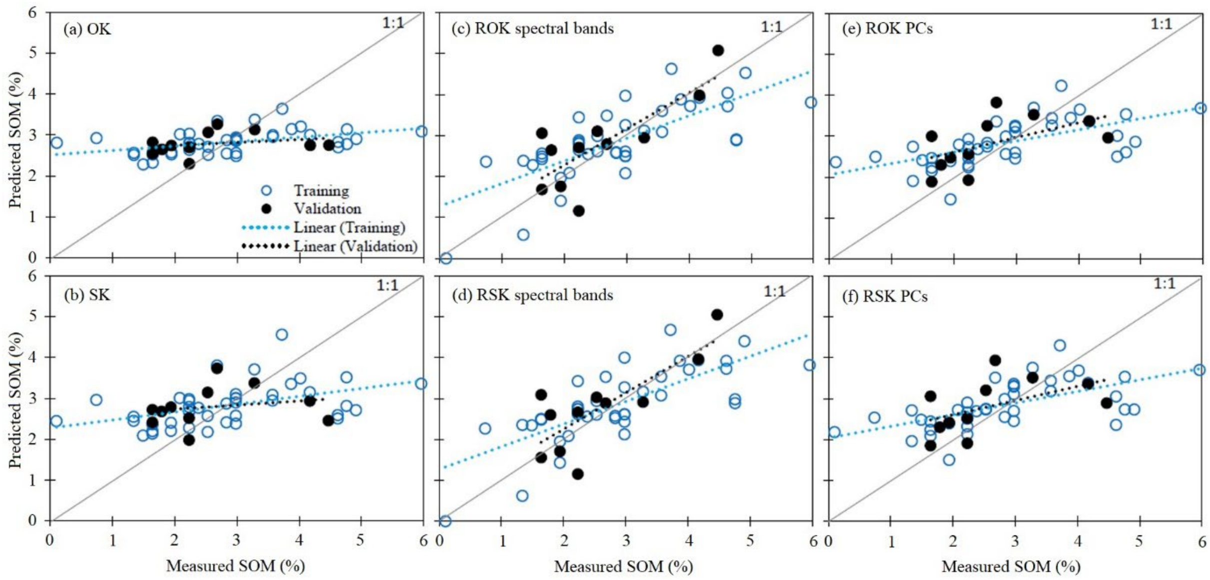

Fig. 2. The observed and predicted SOM content using geostatistical techniques (a) OK, (b) SK; hybrid geostatistical techniques of spectral band (c) ROK, (d) RSK and of PCs (e) ROK and (f) RSK.

Table 2 presents a summary of the model performance for each interpolation method used to predict SOM content. The hybrid methods using spectral bands achieved the highest accuracy (R^2 = 0.63 for both ROK and RSK) with relatively the smallest ME (ROK = 0.53%; RSK = 0.54%) and RMSE (ROK = 0.68%; RSK = 0.67%). It is clear from the ME values that the ROK for the spectral bands had the lowest total ME value (0.53%) and that the RSK had the highest value (0.83%). This was followed by the hybrid methods using PCs, where ROK and RSK, respectively, achieved R^2 values of 0.30 and 0.26. Their error values, respectively, were ME (0.70% and 0.71%) and RMSE (0.82% and 0.85%). Ordinary geostatistical methods applied to the least performed models presented with OK (R^2 = 0.07) and SK (R^2 = 0.04) values, with the highest ME (0.79% and 0.83%) and RMSE (0.93% and 0.98%) values, respectively. Otherwise, all RMSE values appeared to be higher than the ME values.

Table 2. Summary of different interpolation methods based on the validation dataset

Interpolation method	R^2	ME (%)	RMSE (%)
OK	0.07	0.79	0.93
SK	0.04	0.83	0.98
ROK of spectral bands	0.63	0.53	0.68
RSK of spectral bands	0.63	0.54	0.67
ROK of PCs	0.30	0.70	0.82
RSK of PCs	0.26	0.71	0.85

3.2. Geospatial prediction of SOM patterns

The maps in Fig. 3 (a–f) show SOM predictions determined from various geostatistical methods and their variants. Broadly speaking, these maps show similar configurations, with different intensities of SOM content. Except for the central east, the SOM content is obviously larger in the western section of the study area that is dominated by agricultural activities. Noted also, is a remarkably clear declining pattern of SOM content from the central to the southeastern parts of the study area. Otherwise, a patchy distribution of SOM is visible in other parts of this test site.

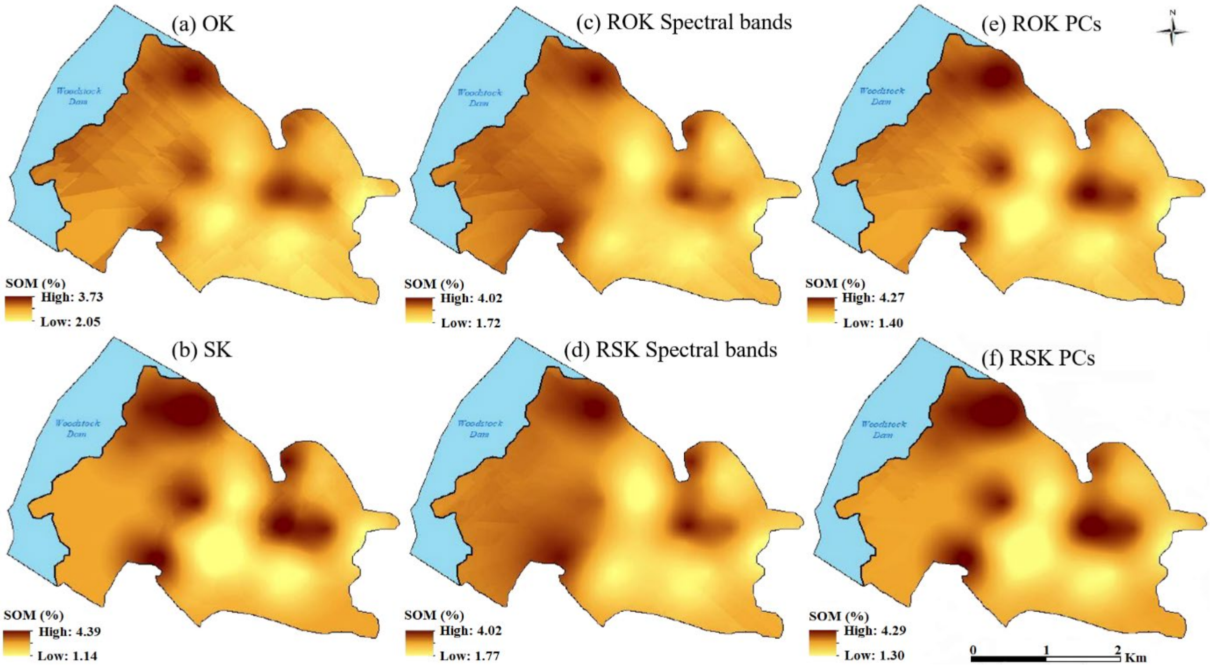

Fig. 3. Geospatial pattern of the predicted SOM content based on geostatistical methods (a) OK and (b) SK; geostatistical varieties with spectral bands (c) ROK, (d) RSK, with PCs (e) ROK and (f) RSK.

4. Discussion

This research was intended to predict the SOM content using various geostatistical methods and their variants. Of the six interpolation techniques used, the ROK and RSK using spectral bands presented with the highest R^2 values (0.63 each), predicting 63% of the SOM content. These results are consistent with existing works (Zhang *et al.*, 2012; Mirzaee *et al.*, 2016a; Song *et al.*, 2017), where RK hybrid geostatistical methods more effectively predicted SOM than ordinary geostatistical methods. This is due to the fact that the hybrid methods are able to incorporate additional information (Hengl, Heuvelink and Stein, 2004). Remote sensing data have been shown to serve as a crucial auxiliary variable to improve estimates of SOM content and its geographic variability (Mirzaee *et al.*, 2016b). The ROK and RSK plots showed a relatively consistent pattern in that they indicated a strong positive linear relationship between measured and predicted SOM content using spectral data. However, the same hybrid methods using PCs showed virtually similar results to the those derived from spectral data, but with a relatively weaker positive linear correlation. According to Mirzaee *et al.* (2016b), there is a non-linear association between the SOM content and the PCs. Consequently, the relationship between the measured and expected SOM is linear and positive, but it is weaker. On the other hand, the results for the ordinary methods, OK and SK, show steady trends, suggesting that the relationship between the measured and predicted SOM content is minimal or non-existent.

Spatially, the SOM content predicted by ordinary and hybrid geostatistical methods shows a relatively similar pattern, with the western part of the study area presenting with the largest concentration. This is attributed to agricultural areas, which also dominate in the western region. This area consists of small-scale farms that play an important role in influencing the formation and maintenance of SOM content (see Fig. 4). For example, since crops are regularly grown in agricultural areas, there is a regular supply of organic matter in the form of leaf litter and stems added to the soil (Behera and Prasad, 2020). Additionally, soils in agricultural landscapes are typically maintained regularly through irrigation and manure applications to increase crop yields. This adds to the enrichment of the SOM content (Bot and Benites, 2005). In part, rangeland could also have contributed to a higher proportion of SOM, as it often extends over large naturally vegetated areas and includes animal manure from livestock grazing. As such, it can influence the SOM content in the region (Haynes, Dominy and Graham, 2003). The central to eastern regions of the study area are subject to severe erosion (Fig. 4(a; b)), which may explain the lower percentages of SOM content. Phuong and Son (2017) noted that SOM and rich topsoil are lost through soil erosion. In their SOM prediction of a mountainous region in China, Liu *et al*. (2015) observed high spatial variability in various land-use classes. Overall, our results highlight the importance of Sentinel-2 imagery and hybrid geostatistical methods for predicting SOM content.

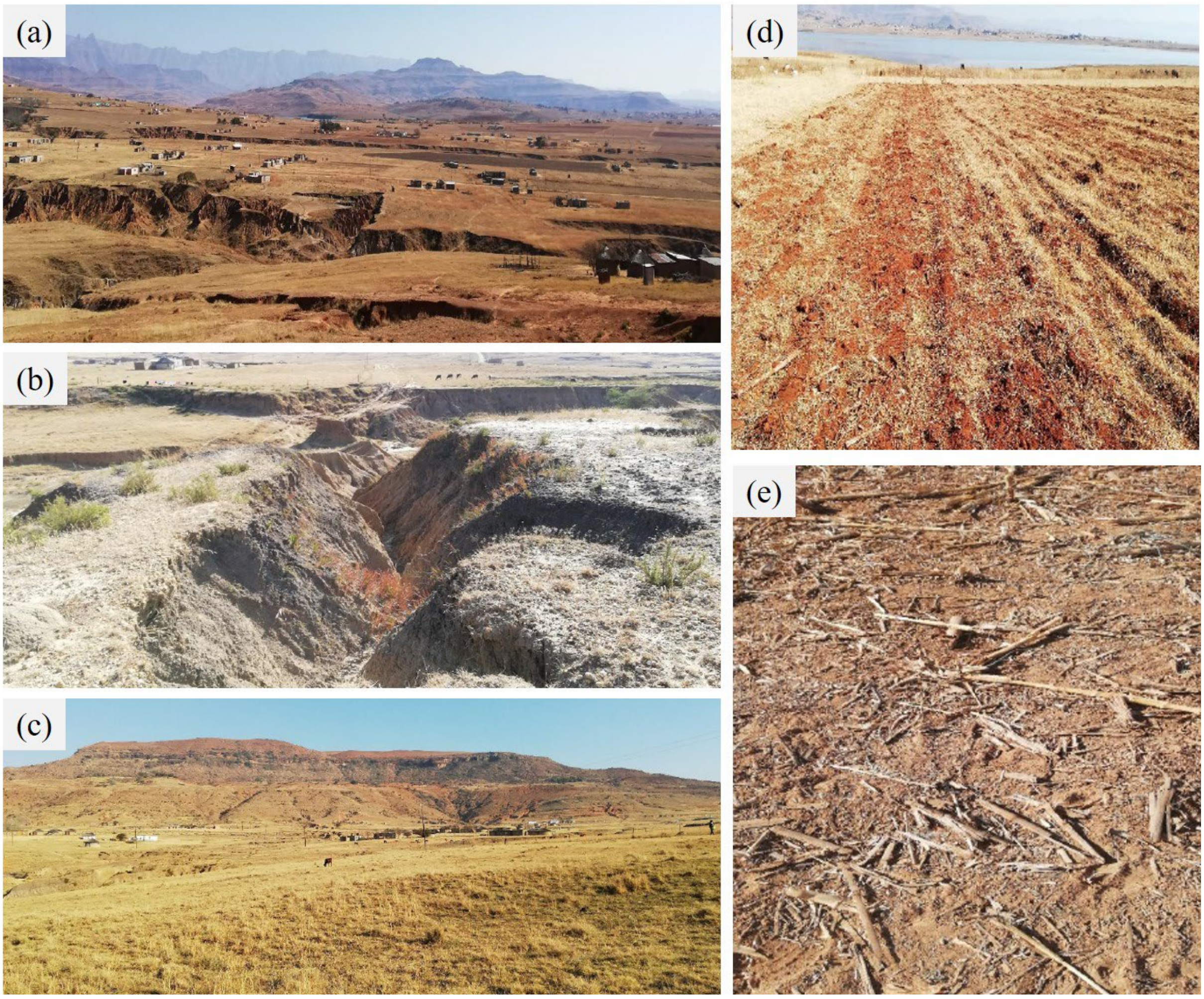

Fig. 4. Landscape conditions in the study area: − severe erosion problems in the central (a) and eastern parts (b); rangeland in the central eastern parts (c); cropland in the fallow season (d); and maize crop residues dispersed across the farmland (e).

5. Limitations

Despite the success of hybrid geostatistical methods for the prediction of SOM, the study is limited in spatial scope. It could have been extended to the uThukela River Catchment. The study also involved a relatively small (52) sample that was split into training and validation. Therefore, cross validation is recommended in the future. Future research should also include spatial autocorrelation and uncertainty analysis for predictions, thereby highlighting the advantages of geostatistical methods over other existing methods.

6. Conclusion

In this study, we evaluated soil organic matter (SOM) predictions using geostatistical and hybrid methods based on Sentinel-2 imagery. We found that hybrid geostatistical techniques (ROK and RSK) outperformed ordinary geostatistical methods (OK and SK). Specifically, we found that by supplementing the models with auxiliary Sentinel-2 data, such as spectral bands and their combinations, it was possible to significantly improve model prediction performance and accuracy. Conversely, the lack of such supplementary data resulted in poor and inaccurate predictions. Overall, ROK and RSK, which used the relevant spectral bands, predicted SOM content most accurately. They were followed by ROK and RSK, which used principal components, and with OK and SK proving to be the least accurate. Therefore, we conclude that incorporating Sentinel-2 spectral data can considerably improve SOM content predictions. We do, however, recommend the testing of other validation methods, such as cross-validation. Because Sentinel-2 imagery is readily available at high spatial and temporal resolutions, the combination of Sentinel-2 and geostatistical methods suggests that it may prove highly useful in predicting SOM content. Using such techniques has potential for the sustainable monitoring of SOM content on varying geographic scales. Furthermore, important conclusions about SOM variability can be derived to guide land management practices, particularly for countries such as South Africa, with a paucity of SOM information, and whose economies depend mainly on land-based resources. Further studies are recommended to quantify and predict SOM content on a broader scale.

7. References

Amulothu, D. *et al.* (2023) ‘Microbial Responses to Carbon Sequestration Soil Amendment and Productivity’, *International Journal of Environment and Climate Change*, 13(10), pp. 3587–3597.

Askari, M.S. *et al.* (2020) ‘Quantification of heavy metal pollution for environmental assessment of soil condition’, *Environmental Monitoring and Assessment*, 192(3), p. 162. Available at: https://doi.org/10.1007/s10661-020-8116-6.

Aubert, P.-M. *et al.* (2020) ‘Holding the ground. Alliances and defiances between scientists, policy-makers and civil society in the development of a voluntary initiative, the “Four per 1000 Soils for Food Security and Climate”’, *Environmental Science & Policy*, 113, pp. 80–87. Available at: https://doi.org/10.1016/j.envsci.2020.06.008.

Bangroo, S.A. *et al.* (2020) ‘Application of predictor variables in spatial quantification of soil organic carbon and total nitrogen using regression kriging in the North Kashmir forested Himalayas’, *Catena*, 193, p. 104632.

Behera, B.K. and Prasad, R. (2020) Environmental technology and sustainability: Physical, chemical and biological technologies for clean environmental management. Elsevier. Available at: https://books.google.co.za/books?hl=en&lr=&id=u0bbDwAAQBAJ&oi=fnd&pg=PP1&dq=Behera,+B.+K.+and+Prasad,+R.+(2020).+Enhance+organic+matter.+Chapter+4+Strategies+for+soil+management.+Environmental+Technology+and+Sustainability:+Physical,+Chemical+and+Biological+Technologies+for+Clean+Environmental+Management.+Elsevier:+153.&ots=Q8pEG7KSTA&sig=VNMVGePYOFtrLE66eDuPtj2GZmQ. (Date accessed: 30 March 2024).

Ben-Dor, E., Irons, J.R. and Epema, G.F. (1999) ‘Soil reflectance’, *Remote Sensing for the Earth Sciences: Manual of Remote Sensing*, 3(3), pp. 111–188.

Bhunia, G.S., Shit, P.K. and Maiti, R. (2018) ‘Comparison of GIS-based interpolation methods for spatial distribution of soil organic carbon (SOC)’, *Journal of the Saudi Society of Agricultural Sciences*, 17(2), pp. 114–126.

Bot, A. and Benites, J. (2005) *The importance of soil organic matter: key to drought-resistant soil and sustained food production*. Food & Agriculture Org. (80). Available at:

https://books.google.co.za/books?hl=en&lr=&id=dJe5-pmmjaAC&oi=fnd&pg=PR8&dq=Bot,+A.+and+Benites,+J.+(2005).+The+importance+of+soil+organic+matter:+Key+to+drought-resistant+soil+and+sustained+food+production.+Food+and+Agriculture+Organisation,+80:+1-95.&ots=FyuYY98ppf&sig=ZWITlF2qVtu3CWNtoBIlOJ8anCk (Date accessed: 17 February 2024).

Castaldi, F. *et al.* (2019) 'Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands', *ISPRS Journal of Photogrammetry and Remote Sensing*, 147, pp. 267–282.

Chang, N. and Chen, D. (2024) 'Prediction of soil organic matter using Landsat 8 data and machine learning algorithms in typical karst cropland in China', *European Journal of Agronomy*, 160, p.127323.

Craswell, E.T. and Lefroy, R.D.B. (2001) 'The role and function of organic matter in tropical soils', 61(1), pp. 7–18.

Drusch, M. *et al.* (2012) 'Sentinel-2: ESA's optical high-resolution mission for GMES operational services', *Remote Sensing of Environment*, 120, pp. 25–36.Du Preez, C.C., van Huyssteen, C.W. and Mnkeni, P.N. (2011) 'Land use and soil organic matter in South Africa 1: A review on spatial variability and the influence of rangeland stock production.', *South African Journal of Science*, 107. Available at: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=00382353&AN=69815298&h=sIAZ5BH7op3yvj6AxOy07Qm7I3vqkjyOqmVELj5VOzr4njLH8GrQHdH4IHj7QX8nsyAG331R%2FTW3n1STqj0rqA%3D%3D&crl=c (Date accessed: 30 March 2024).

ESA (2015) 'Sentinel-2 User Handbook'. European Space Agency. Available at: https://earth.esa.int/documents/247904/685211/Sentinel-2_User_Handbook (Date accessed: 7 October 2023).

ESA (2020) *Sentinel-2 MSI User Guide - Product Types*. Available at: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/product-types (Date accessed: 10 July 2023).

Gallo, B.C. *et al.* (2018a) 'Multi-temporal satellite images on topsoil attribute quantification and the relationship with soil classes and geology', *Remote Sensing*, 10(10), p. 1571.

Gallo, B.C. *et al.* (2018b) 'Multi-temporal satellite images on topsoil attribute quantification and the relationship with soil classes and geology', *Remote Sensing*, 10(10), p. 1571.

Gattinger, A. *et al.* (2011) 'No-till agriculture – a climate-smart solution?' Available at: https://orgprints.org/20302/ (Date accessed: 30 March 2024).

Gholizadeh, A. *et al.* (2018) 'Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging', *Remote Sensing of Environment*, 218, pp. 89–103.

Gia Pham, T. *et al.* (2019) 'Application of ordinary kriging and regression kriging method for soil properties mapping in hilly region of Central Vietnam', *ISPRS International Journal of Geo-Information*, 8(3), p. 147.

Guo, L. *et al.* (2018) 'Spatial modelling of soil organic carbon stocks with combined principal component analysis and geographically weighted regression', *The Journal of Agricultural Science*, 156(6), pp. 774–784.

Haynes, R.J., Dominy, C.S. and Graham, M.H. (2003) 'Effect of agricultural land use on soil organic matter status and the composition of earthworm communities in KwaZulu-Natal, South Africa', *Agriculture, Ecosystems & Environment*, 95(2–3), pp. 453–464.

Hengl, T., Heuvelink, G.B. and Rossiter, D.G. (2007) 'About regression-kriging: From equations to case studies', *Computers & Geosciences*, 33(10), pp. 1301–1315.

Hengl, T., Heuvelink, G.B. and Stein, A. (2004) 'A generic framework for spatial prediction of soil variables based on regression-kriging', *Geoderma*, 120(1–2), pp. 75–93.

Juma, N.G. (1999) *The pedosphere and its dynamics: a systems approach to soil science*. Salman Productions.

King, J.K.K. *et al.* (2018) 'Soil sciences and the French "Four per 1000" Initiative—The promises of underground carbon', *Energy Research & Social Science*, 45, pp. 144–152.

Krishan, G. *et al.* (2014) 'Remote sensing in soil fertility evaluation and management', *Bioresources for sustainable plant nutrient management. New Delhi: Satish Serial Publishing House*, pp. 509–533.

Kumar, N. *et al.* (2018) 'Geospatial Mapping of Soil Organic Carbon using Regression Kriging and Remote Sensing', *Journal of the Indian Society of Remote Sensing*, 46(5), pp. 705–716. Available at: https://doi.org/10.1007/s12524-017-0738-y.

Lal, R. (2020a) 'Food security impacts of the "Four per Thousand" Initiative', *Geoderma*, 374, p. 114427.

Lal, R. (2020b) 'The role of industry and the private sector in promoting the "Four per 1000" Initiative and other negative emission technologies', *Geoderma*, 378, p. 114613. Available at: https://doi.org/10.1016/j.geoderma.2020.114613.

Li, H. *et al.* (2024) 'Soil organic matter content prediction based on two-branch convolutional neural network combining image and spectral features', *Computers and Electronics in Agriculture*, 217, p. 108561.

Li, X. *et al.* (2021) 'Digital mapping of soil organic carbon using Sentinel series data: A case study of the Ebinur Lake Watershed in Xinjiang', *Remote Sensing*, 13(4), p. 769.

Liu, H., Zhang, Y. and Zhang, B. (2009) 'Novel hyperspectral reflectance models for estimating black-soil organic matter in Northeast China', *Environmental Monitoring and Assessment*, 154(1–4), pp. 147–154. Available at: https://doi.org/10.1007/s10661-008-0385-4.

Liu, S. *et al.* (2015) 'Prediction of soil organic matter variability associated with different land use types in mountainous landscape in southwestern Yunnan province, China', *Catena*, 133, pp. 137–144.

Lorenz, K. and Lal, R. (2018) 'Agricultural Land Use and the Global Carbon Cycle', in Lorenz, K. and Lal, R., *Carbon Sequestration in Agricultural Ecosystems*. Cham: Springer International Publishing, pp. 1–37. Available at: https://doi.org/10.1007/978-3-319-92318-5_1.

McBratney, A.B., Santos, M.M. and Minasny, B. (2003) 'On digital soil mapping', *Geoderma*, 117(1–2), pp. 3–52.

Meng, X. *et al.* (2020) 'Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data', *International Journal of Applied Earth Observation and Geoinformation*, 89, p. 102111.

Mirzaee, S. *et al.* (2016a) 'Spatial variability of soil organic matter using remote sensing data', *Catena*, 145, pp. 118–127.

Mirzaee, S. *et al.* (2016b) 'Spatial variability of soil organic matter using remote sensing data', *Catena*, 145, pp. 118–127.

Nehrani, S.H., Askari, M.S., Saadat, S., Delavar, M.A., Taheri, M. and Holden, N.M. (2020) 'Quantification of soil quality under semi-arid agriculture in the northwest of Iran', *Ecological Indicators*, 108, p.105770.

Phuong, T.T. and Son, N.T. (2017) ‘Land use change and its interactions with soil, water resources, and rural livelihoods in Hoa Binh province’, *Vietnam Journal of Agricultural Sciences*, 15(3), pp. 249–262.

Piccini, C., Marchetti, A. and Francaviglia, R. (2014) ‘Estimation of soil organic matter by geostatistical methods: Use of auxiliary information in agricultural and environmental assessment’, *Ecological Indicators*, 36, pp. 301–314.

Pouladi, N. *et al.* (2023) ‘Digital mapping of soil organic carbon using remote sensing data: A systematic review’, *Catena*, 232, p. 107409. Available at: https://doi.org/10.1016/j.catena.2023.107409.

Robinson, T.P. and Metternicht, G. (2006) ‘Testing the performance of spatial interpolation techniques for mapping soil properties’, *Computers and Electronics in Agriculture*, 50(2), pp. 97–108.

Roy, D.P. *et al.* (2016) ‘Best practices for the reprojection and resampling of Sentinel-2 Multi Spectral Instrument Level 1C data’, *Remote Sensing Letters*, 7(11), pp. 1023–1032. Available at: https://doi.org/10.1080/2150704X.2016.1212419.

Samuels, P. (2017) ‘Advice on exploratory factor analysis’. Available at: https://www.open-access.bcu.ac.uk/6076/ (Date accessed: 17 February 2024).

Schulze, R.E. (1997) *South African Atlas of agrohydrology and climatology: Contribution towards a final report to the Water Research Commission on Project 492: modelling impacts of the Agricultural Environment on Water Resources.* Pietermaritzburg: Water Research Commission (WRC).

Schütte, S., Schulze, R.E. and Paterson, G. (2019) ‘Identification and mapping of soils rich in organic carbon in South Africa as a climate change mitigation option’, *Department of Environmental Affairs, Pretoria, South Africa* [Preprint].

Sewpersad, T., Xulu, S. and Gebreslasie, M. (2024) ‘The assessment of soil organic matter in the KwaZulu-Natal province of South Africa and its relationship to spectroscopy data’, *Geocarto International*, 39(1), p.2361702.

Sharma, V., Chauhan, R. and Kumar, R. (2021) ‘Spectral characteristics of organic soil matter: A comprehensive review’, *Microchemical Journal*, 171, p. 106836.

Shepherd, K.D. and Walsh, M.G. (2007) ‘Infrared spectroscopy − enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries’, *Journal of Near Infrared Spectroscopy*, 15(1), pp. 1–19.

Smith, P. *et al.* (2018) ‘The changing faces of soil organic matter research’, *European Journal of Soil Science*, 69(1), pp. 23–30. Available at: https://doi.org/10.1111/ejss.12500.

Song, Y.-Q. *et al.* (2017) ‘Spatial prediction of soil organic matter using a hybrid geostatistical model of an extreme learning machine and ordinary kriging’, *Sustainability*, 9(5), p. 754.

Statistics South Africa (SSA) (2011) *Statistics South Africa*. Pretoria: Statistics South Africa.

Stevens, A. and Ramirez-Lopez, L. (2014) ‘An introduction to the prospector package’. February. Available at: ftp://journal.r-project.org/pub/R/web/packages/prospectr/vignettes/prospectr.html (Date accessed: 30 March 2024).

Suleymanov, A. *et al.* (2024) ‘Soil organic carbon stock retrieval from Sentinel-2A using a hybrid approach’, *Environmental Monitoring and Assessment*, 196(1), p. 23. Available at: https://doi.org/10.1007/s10661-023-12172-y.

Verster, P.S.J. (1998) *Geological map of Harrismith (2828)*. Pretoria: Council for Geoscience. Available at: https://scholar.google.com/scholar_lookup?title=Geological%20Map%20of%20Harrismith%20&publication_year=1998&author=P.S.J.%20Verster (Date accessed: 17 February 2024).

Viscarra Rossel, R. *et al.* (2006) 'Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties', *Geoderma*, 131(1–2), pp. 59–75.

Wang, J. *et al.* (2017) 'Quantitative estimation of organic matter content in arid soil using Vis-NIR spectroscopy preprocessed by fractional derivative', *Journal of Spectroscopy*, 2017. Available at: https://www.hindawi.com/journals/jspec/2017/1375158/abs/ (Date accessed: 30 March 2024).

Wang, Z., Wu, W. and Liu, H. (2024) 'Spatial estimation of soil organic carbon content utilizing PlanetScope, Sentinel-2, and Sentinel-1 data', *Remote Sensing*, 16(17), p.3268.

Webster, R. and Oliver, M.A. (2007) *Geostatistics for environmental scientists*. John Wiley & Sons. Available at: https://books.google.co.za/books?hl=en&lr=&id=WBwSyvIvNY8C&oi=fnd&pg=PR5&dq=Webster,+R.+and+Oliver,+M.A.+(2001).+Geostatistics+for+environmental+scientists.+John+Wiley+%26+Sons:+183-184.&ots=CEInRUqNZ9&sig=MS09rx7YGro4nFVrKkZYo-mk_Zw (Date accessed: 30 March 2024).

Yao, X. *et al.* (2013) 'Comparison of four spatial interpolation methods for estimating soil moisture in a complex terrain catchment', *PloS One*, 8(1), p. e54660.

Zhang, Meiwei *et al.* (2021) 'Mapping regional soil organic matter based on Sentinel-2a and modis imagery using machine learning algorithms and Google Earth engine', *Remote Sensing*, 13(15), p. 2934.

Zhang, S. *et al.* (2012) 'Spatial prediction of soil organic matter using terrain indices and categorical variables as auxiliary information', *Geoderma*, 171, pp. 35–43.

Zheng, G. *et al.* (2016) 'Estimation of organic matter content in coastal soil using reflectance spectroscopy', *Pedosphere*, 26(1), pp. 130–136.

Žížala, D., Minařík, R. and Zádorová, T. (2019) 'Soil organic carbon mapping using multispectral remote sensing data: Prediction ability of data with different spatial and spectral resolutions', *Remote Sensing*, 11(24), p. 2947.