Abstract: Cation exchange capacity is a crucial metric for managing soil fertility and promoting agricultural sustainability. An alternative technique for the non-destructive assessment of important soil parameters is reflectance spectroscopy. The main focus of this paper is on how to analyze and predict the content of various soil cation exchange capacities (CEC) in arid conditions (Sohag governorate, Egypt) at a low cost using laboratory analysis of CEC, visible near-infrared and shortwave infrared (Vis-NIR) spectroscopy, partial least-squares regression (PLSR), and Ordinary Kriging (OK). Utilizing reflectance spectroscopy with a spectral resolution of 10 nm and laboratory studies with a spectral range of 350 to 2500 nm, 104 surface soil samples were collected to a depth of 30 cm in the Sohag governorate, Egypt (which is part of the dry region of North Africa), in order to accomplish this goal. The association between the spectroradiometer and CEC averaged values was modeled using PLSR in order to map the predicted value using Ordinary Kriging (OK). Thirty-one soil samples were selected for validation. The predictive validity of the cross-validated models was evaluated using the coefficient of determination ($R^2$), root mean square error (RMSE), residual prediction deviation (RPD), and ratio of performance to interquartile distance (RPIQ). The results indicate that ten transformation methods yielded calibration models that met the study’s requirements, with $R^2 > 0.6$, RPD > 2.5, and $R_{iQP} > 4.05$. For evaluating CEC in Vis-NIR spectra, the most efficient transformation and calibration model was the reciprocal of Log R transformation ($R^2 = 0.98$, RMSE = 0.40, RPD = 6.99, and $R_{iQP} = 9.22$). This implies that combining the reciprocal of Log R with PLSR yields the optimal model for predicting CEC values. The CEC values were best fitted by four models: spherical, exponential, Gaussian, and circular. The methodology used here does offer a “quick”, inexpensive tool that can be broadly and quickly used, and it can be readily implemented again in comparable conditions in arid regions.

Keywords: predicative model; PLSR; CEC; soil fertility; Vis-NIR; agricultural sustainability; drylands

1. Introduction

Managing soil fertility and advancing agricultural sustainability require a measurement known as cation exchange capacity (CEC). Traditionally, components such as fertility have been determined through laboratory soil analysis. This is time-consuming and expensive, which raises concerns for precision agriculture’s future, which has stalled in several nations [1]. The ability of soil to adsorb and exchange cations is known as its cation exchange capacity (CEC). Furthermore, CEC is a crucial measure for mineralogical characterization and soil fertility [2,3]. One of the primary chemical characteristics of soil is...
its cation exchange capacity (CEC), which is necessary for several disciplines, including soil science, agricultural and environmental engineering, and soil science [4].

For the determination of CEC, a number of traditional laboratory analysis techniques have historically been available [5]. According to Richards [6] the majority of procedures entail saturating the soil with an index cation (NH$_4$), eliminating surplus cations by washing, replacing the adsorbed index cation with another cation (Na), and measuring the index cation in the final extract. Furthermore, due to increased Ca solubility in calcareous and gypsiferous soils, new approaches have been implemented [7]. However, the current methods are costly, labor-intensive, and time-consuming, especially when dealing with a lot of soil samples [8,9]. The scientific community is now working to provide quick and affordable ways for evaluating particular factors because of the difficulties involved with the traditional approaches. Although there has been significant advancement in the accurate application of machine learning algorithms for CEC prediction, these algorithms have not yet been applied to a variety of land uses worldwide [10].

One of the most promising methods for making fast decisions on soil qualities is chemometric analysis, which makes use of the visible, near-infrared, and shortwave infrared reflectance domains to predict different soil parameters. Vis-NIR spectroscopy has been shown in several studies to be useful for estimating soil properties, including CEC [11–13]. Through the use of Vis-NIR spectra, information is mostly connected to the organic matter (OM) content, water content, and mineral elements of the soil, such as clay minerals [14]. This information is encoded in the spectrum by its absorptions (spectral characteristics) at particular wavelengths, intensities, and shapes [15]. Additionally, the Vis-NIR approach enables the calculation of secondary soil parameters, or characteristics like pH, cation exchange capacity (CEC), base saturation (V), and exchangeable (ex-) nutrients that have indirect correlations with the spectrum [16]. When primary and secondary soil properties covariate, this is achievable [17]. The method’s ability to complete quick, non-destructive analysis with little to no sample preparation has led to its increasing use. It overcomes the drawbacks of wet chemistry by offering a chemical-free, real-time evaluation of soil quality that may be applied directly at a field location. In reaction to an infrared (IR) source, it creates a fingerprint of the sample that is closely related to its physicochemical makeup [18]. The rapid predictability of this technique when combined with statistical multivariate modeling makes it a viable alternative that has gained popularity for a number of nutrients, including soil organic carbon, nitrogen, phosphorus, potassium, and other nutrients [19–21].

The partial least-squares regression (PLSR) technique is the most well-known and extensively utilized multivariate statistical technique in chemometric research, with researchers worldwide using it [22]. PLSR is very useful when there is multicollinearity among predictors (X variables) and when the predictor matrix has many more variables than the number of observations. Because of its greater understandability and faster processing algorithm, PLSR is typically chosen over other multivariate statistical techniques [23]. Numerous researchers, such as Mousavi et al. [24] and Yitagesu et al. [25], have revealed that the primary benefit of PLSR is its great efficiency when dealing with small sample sizes.

Soil spectral data have been preprocessed using a variety of techniques to highlight characteristics, reduce noise, and prepare them for estimation modeling. For example, to find the optimal combination for estimating soil carbon, ref. [26] employed five multivariate approaches and thirty preprocessing transformations, such as derivatives, normalization, and non-linear transformations. The best performance, according to the authors, was shown by PLSR, which was followed by stepwise multiple linear regression (SMLR). Among the preprocessing adjustments, Savitzky–Golay derivatives often improved models for all multivariate approaches. Additionally, five machine learning (ML) model predictions and spatial uncertainties were developed and thoroughly assessed from the standpoint of a soil scientist by Kaya et al. [27]. Higher $R^2$ (0.41) values and lower RMSE (4.12 cmolc kg$^{-1}$) and MAE (3.12 cmolc kg$^{-1}$) values showed that the RF model performed better than
the other models. Several artificial intelligence (AI) methods were employed by other studies to predict CEC [28,29]. Today, visible near-infrared and shortwave infrared (Vis-NIR) approaches are generally acknowledged to be useful for quantitatively retrieving soil parameters and typically yield good indicators of soil quality [30]. However, a number of important issues remain to be resolved, including the proportion of gypsum and carbonate, which is still a contentious subject today. Appropriate multivariate regression techniques, which have an advantage over some bivariate associations and are appropriate for peak intensity data, can be used to obtain quantitative information about CEC. Additionally, a variety of multivariate methods such as main components regression, multiple linear regression, and partial least squares stepwise linear regression (PLSR) support vector machines and multivariate adaptive regression splines can be utilized for several kinds of titrated soil examination [31]. With all of the above-mentioned details in mind, the current study set out to create predictive modeling of soil CEC in arid conditions (Sohag governorate, Egypt) utilizing a variety of spectral transformations and the PLSR approach.

2. Materials and Methods

2.1. Experimental Area

The study area covers a part (alluvial plain about 2021.1 km²) of Sohag governorate, Egypt, which extends from the northern edge of Qena Governorate at latitude 26°07’N to the southern edge of Assiut Governorate at latitude 26°57’N. It is bounded between longitudes 31°20’E and 32°14’E with elevation varying from about 62 to 75 m above the mean sea level (Figure 1). The study area belongs to the arid region of North Africa, which is generally characterized by hot summers and mild winters with low rainfall. Air temperature ranges between 36.5 °C (summer) and 15.5 °C (winter), relative humidity ranges between 51% and 61% (winter), 33% and 41% (spring), and 35% and 42% (summer) ([32], Rainfall is generally rare and randomly precipitates over the area. The Nile River, irrigation canals, and drainage systems hydrologically reflect the surface water in the Sohag region. The Nile’s alluvial deposits created the Quaternary aquifer system in the research region, which is divided into two strata with different hydraulic characteristics. The clay–silt component, which is the top layer, has little permeability both vertically and horizontally. The primary aquifer with high horizontal and vertical permeability is formed by the lowest layer, the graded sand member [33].

![Figure 1. Location map of the studied area.](image)

2.2. Field Work and Lab Analysis

A total of 104 representative sites were chosen to encompass the study region (Figure 2). A hand-held GPS (Germin e-trix, Nevada, KS, USA) was used to record the location coordinates using the WGS 84 coordinate system. The sampling sites were chosen based on the spatial heterogeneity seen in the alluvial plain. A random sampling method was utilized. This entailed taking soil samples at random sites throughout the selected sampling
region. Each soil sample was taken at each sampling location from three to five disturbed
subsamples in the surface layer (0 to 30 cm); the soils were then combined and put into
polyethylene bags for additional examination. The soil samples were allowed to air dry
at room temperature; consequently, the dried soil samples were crushed using a wooden
mortar and then sieved through a 2 mm sieve. The salinity (electric conductivity) with an
EC meter (Orion, Model 150, Bosque Farms, NM, USA) [34], soil acidity (pH) in a saturated
paste with a pH meter supplied from HANNA Instruments (Leighton Buzzard, UK) [35],
and cation exchange capacity (CEC) [36] were measured. Furthermore, in accordance
with [37], the particle size distribution was measured.

A spectroradiometer FieldSpec 3 (ASD-Analytical Spectral Devices, Boulder, CO, USA)
with a spectral range of 350 to 2500 nm and a spectral resolution of 10 nm, with a sampling
interval of 1 nm, was used to collect the Vis-NIR laboratory-based spectral measurements
under standard illumination conditions using two calibrated tungsten quartz halogen lamps
in the dark. Petri dishes with roughly 95 cm$^3$ of soil were filled with soil samples. For every
soil sample, the equipment calculated an average of 100 scans. The spectroradiometer was
calibrated using a white Spectralon with 100% reflectance (LabSphere, North Sutton, NH,
USA) prior to scanning [38].

Figure 2. Soil sample locations over the study area.

2.3. Acquisition of Spectral Data

A spectroradiometer FieldSpec 3 (ASD-Analytical Spectral Devices, Boulder, CO, USA)
with a spectral range of 350 to 2500 nm and a spectral resolution of 10 nm, with a sampling
interval of 1 nm, was used to collect the Vis-NIR laboratory-based spectral measurements
under standard illumination conditions using two calibrated tungsten quartz halogen lamps
in the dark. Petri dishes with roughly 95 cm$^3$ of soil were filled with soil samples. For every
soil sample, the equipment calculated an average of 100 scans. The spectroradiometer was
calibrated using a white Spectralon with 100% reflectance (LabSphere, North Sutton, NH,
USA) prior to scanning [38].

2.4. Data Preparation and Preprocessing Transformation

Utilizing a moving Savitzky–Golay filter algorithm [39] with a window width of 5 nm
and a second-degree polynomial, the instrument noise was eliminated. The number of
predictor variables in the soil Vis-NIR spectra was lowered from 2150 to 1075 by first
averaging the data over two bands. This shortened the dataset’s dimensionality, slashed
processing time, and made them ready for additional preparation. Preprocessing is often
the first step in spectroscopic analysis, with the goal of assessing and potentially improving
data quality. The spectral preprocessing method uses many mathematical strategies to
improve the data and rectify light scattering in reflectance measurements before using the
data in calibration models [40]. A comprehensive literature review was used to identify
several reflectance transformations that are responsive to diverse soil conditions in the
current investigation. The authors of [22,23,41,42] have provided a full description of the
development of various reflectance transformations and their usage in the quantitative
modeling of various soil parameters. In the current investigation, 18 distinct reflectance
transformations were identified based on all of those studies, and their potential for quan-
titative CEC prediction was assessed using these transformations, i.e., Reflectance (R),
Logarithm of R (Log R), Square root of R (sqrt R), Reciprocal of R (1/R), Reciprocal of Log R
(1/Log R), Reciprocal of sqrt R (1/sqrt R), First derivative of R (R'), First derivative of Log R
(Log R'), First derivative of sqrt R (sqrt R'), Second derivative of R (R''), Second derivative of Log R
(Log R''), Second derivative of sqrt R (sqrt R''), First derivative of 1/R (1/R'), First
derivative of 1/Log R (1/Log R'), First derivative of 1/sqrt R (1/sqrt R'), Second derivative of
1/R (1/R''), Second derivative of 1/Log R (1/Log R'')→, and Second derivative of 1/sqrt R
(1/sqrt R''). The production of reflectance changes was performed using the spectral libraries
that were created for each sample. Various reflectance value transformations help with
hyperspectral modeling for the prediction of soil properties by eliminating noise (linear,
linear-like, and multiplication noise), eliminating the additive baseline (or “offset”), and
linearizing the previously non-linear correlation between measured reflectance and soil
physico-chemical properties [23].

2.5. Multiple Variable Statistical Modeling

Using a variety of reflectance modifications, the Partial Least Square Regression
(PLSR) technique was employed to predict soil attributes among the many multivariate
statistical methods available for chemometric analysis. A popular Vis-NIR calibration
method noted for its predictability, simplicity, accuracy, and quantitative explanations is
called PLSR [43,44]. In general, PLSR is preferred due to its faster computational speed and
easier-to-understand algorithm. Essentially, PLSR looks for a small number of components
or factors (linear combinations) within the spectral data (original X values) and uses them
to create regression equations that predict the observed variable (soil properties). Using
laboratory soil data, Vis-NIR spectral data, the R studio software 4.1.2 PLS package was
used to create the various soil parameters’ calibration and validation models in accordance
with the protocol provided by [45]. The orthogonal components that increase the variance
of the predictor (X spectra that are mean-centered before decomposition) and response
variables (Y lab data from the chemical analysis) are chosen using the PLSR algorithm.
Factor loading (P and q) and factor scores (T) are the breakdowns of X and Y provided by
PLSR. Equations (1) and (2) include residues E and F to account for the remaining noise
elements that can be ignored.

\[ X = TP + E \]  \hspace{1cm} (1)
\[ Y = Tq + F \]  \hspace{1cm} (2)

The entire dataset of 104 observations was randomly divided into two groups in
order to determine the accuracy of the PLSR regression models: the validation dataset,
which included 31 samples (30% of the total dataset), was used to test the developed
model’s accuracy, and the calibration dataset, which included 73 samples (70% of the total
dataset), for the development of the prediction model. To evaluate each PLSR component’s
predictive importance, random cross-validation was used.

2.6. Assessment of the Developed Models’ Performance

The coefficient of determination (R^2), root mean square error (RMSE), residual prediction
deviation (RPD), and ratio of performance to interquartile distance (RPIQ) were used
to assess the cross-validated models’ predictive quality as given in Equations (3)–(6).

\[ R^2 = 1 - \frac{\sum(Y_{pred} - Y_{meas})^2}{\sum(Y_i - Y_{meas})^2} \]  \hspace{1cm} (3)
where \( Y_{\text{pred}} \) is the soil predicted values; \( Y_i \) is the soil measured values mean; \( Y_{\text{Meas}} \) is the soil measured values; and \( n \) is the number of measured or predicted values.

\[
RMSE = \sqrt{\frac{1}{n} \sum (Y - \bar{X})^2}
\]  

(4)

where \( Y \) is the soil predicted values; \( \bar{X} \) is the soil measured values; and \( n \) is the number of measured or predicted values.

\[
RPD = \frac{SD}{RMSE}
\]

(5)

where \( SD \) is the standard deviation.

The ratio of performance to IQ (RPIQ) value is more objective than the RPD value; however, Bellon-Maurel noted that both the chemical content and the physical characteristics of the soil show a biased normal distribution [46].

\[
RPIQ = \frac{\text{IQ}_{RMSE} = Q_3 - Q_1}{RMSE}
\]

(6)

where:

The difference between the first quartile (\( Q_1 \)) and the third quartile (\( Q_3 \)), which is where we find 75% of the samples, is known as IQ. \( Q_1 \) is the value below which we can find 25% of the samples.

Regression models of the spectrum have been extensively trained using these methods [47].

The best cross-validated models were those with the lowest RMSE and highest \( R^2 \), RPD, and RPIQ values. The validation dataset was later used to independently test the selected models, and \( R^2 \), RPD, RPIQ, and RMSE values were used to assess the overall prediction accuracy [23]. According to [48], the \( R^2 \) value indicated whether the model was: (1) fairly predicted (\( R^2: 0.50–0.75 \)); (2) unreliable (\( R^2 < 0.50 \)); or (3) well suited to predict the attribute reliably (\( R^2 > 0.75 \)). According to the values of (RPD), six categories of models/predictions were defined [44] as shown in (Table 1).

<table>
<thead>
<tr>
<th>RPD</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1.0</td>
<td>Very poor</td>
</tr>
<tr>
<td>1.0–1.4</td>
<td>Poor</td>
</tr>
<tr>
<td>1.4–1.8</td>
<td>Fair</td>
</tr>
<tr>
<td>1.8–2.0</td>
<td>Good</td>
</tr>
<tr>
<td>2.0–2.5</td>
<td>Very good</td>
</tr>
<tr>
<td>&gt;2.5</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

Whereas, four classes of models were recognized based on RPIQ values [46] and presented in (Table 2).

<table>
<thead>
<tr>
<th>RPIQ</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.02–2.7</td>
<td>Poor</td>
</tr>
<tr>
<td>2.7–3.37</td>
<td>Fair</td>
</tr>
<tr>
<td>3.27–4.05</td>
<td>Good</td>
</tr>
<tr>
<td>&gt;4.05</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

### 2.7. Semivariogram Analysis

A semivariogram is a statistical tool that evaluates how much two random variables’ average degree of similarity decreases as their distance from one another rises. According to [49], it shows the measured sample points’ spatial autocorrelation. The predicted values from the PLSR models were interpolated using ordinary kriging (OK). The value
of the soil parameters at an unsampled location was ascertained using a set of analytical techniques using the OK geostatistics framework. The spatial continuity between points is characterized by the semivariance function. A semivariogram is created when the semivariance is plotted against the lag distance or separation distance between the dots.

Lark [50]. Nugget semivariance, range, and sill were found to be the model’s parameters (Figure 3). Range is defined as the distance beyond which observations are not spatially dependent [51]; sill is the lag distance between measurements at which one value for a variable does not influence neighboring values; and nugget semivariance is the variance at zero distance and represents the experimental error and field variation within the minimum sampling spacing, as seen in Abuzaid [52].

![Figure 3. The relationship between variogram parameters (nugget, partial sill, sill) and sample separation distance.](image)

The spatial dependency (SPD) is defined by the hint(C0)/sill(C0 + C) ratio, where ratios less than 0.25, 0.25–0.75, and above 0.75 indicate a strong, moderate, and weak SPD, respectively [53,54]. To ensure the models’ validity and to compare the estimated values from the semivariogram model with the actual values, the models underwent cross-validation (a process including trial and error). The cross-validation statistics mean error (ME), mean standardized error (MSE), and root mean square standardized error (RMSSE) are used to quantify differences between estimated and experimental values [55]. The semivariogram models’ parameters were cross-validated and then utilized to create maps for CEC using ordinary kriging interpolation in ArcGIS 10.4. The kriging method was applied using the following Equation (7), which was provided by [56]. Based on the lowest ME and MSE (around zero), comparable values of RMSE and ASE, and RMSSE close to unity, the best-fit model was chosen [57,58].

\[
Z'(X_o) = \sum_{i=1}^{N} \lambda_i Z'(x_i)
\]

where \(Z'(x_o)\) is the estimated variable at location \(x_o\), \(Z'(x_i)\) is the value of an inspected variable at location \(x_i\), \(N\) is the number of observations in the vicinity of the inspected point, and \(\lambda_i\) is the statistical weight assigned to \(Z'(x_i)\) for a sample near \(x_o\).

3. Results and Discussion

3.1. Descriptive Statistics of Soil Samples

Table 3 presents an overview of the descriptive statistical analysis of the soil parameters that were examined. The soils ranged in pH from a low of 7.73 to a maximum of 8.48. This shows that the study area’s conditions are slightly to highly alkaline [59]. The pH of the soil greatly affects the nutrients that are available to plants. Iron, zinc, and other
micronutrients are less available to plants when the pH of the soil is too high [60]. As a result, sustaining an appropriate pH range in the soil is essential for plant growth and nutrient availability [60]. Furthermore, increased soil pH causes a rise in the mineralizable portions of N and C ratios, where the connections between organic elements and clays are broken [61]. According to the data, the soils in the research area have slight to high salinity, with average EC values of 4.01 ± 2.75 dSm⁻¹ and a range of 0.80 to 11 dSm⁻¹ [62]. The amounts of sand, silt, and clay in the soil range from 5.05 to 74.34%, 9.07 to 60.91%, and 13.56 to 49.05%, respectively. There were differences in soil textures between loam, clay, clay loam, silty loam, silty clay loam, sandy loam, sandy clay loam, sandy clay, and silty clay (Figure 4). These soils’ cation exchange capacity is moderate, with a mean value of 22.72 cmolc kg⁻¹. The CEC in the studied area ranges widely, from 17.91 to 30.35 cmol c kg⁻¹ soil, with an average of 22.72 ± 2.66 cmol kg⁻¹ soils. The high values of CEC and basic alkali cation adsorption are correlated with clay content [63]. The interpolation maps of EC and pH values are given in Figure 4a,b.

Table 3. Summary of descriptive statistical analysis of the investigated soil parameters.

<table>
<thead>
<tr>
<th>Properties</th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand (%)</td>
<td>40.27</td>
<td>25.05</td>
<td>74.34</td>
<td>9.14</td>
</tr>
<tr>
<td>Silt (%)</td>
<td>33.80</td>
<td>22.11</td>
<td>60.91</td>
<td>6.21</td>
</tr>
<tr>
<td>Clay (%)</td>
<td>25.93</td>
<td>13.56</td>
<td>49.05</td>
<td>8.07</td>
</tr>
<tr>
<td>CEC (cmolc kg⁻¹)</td>
<td>22.72</td>
<td>17.91</td>
<td>30.35</td>
<td>2.66</td>
</tr>
<tr>
<td>EC (dSm⁻¹)</td>
<td>4.01</td>
<td>0.80</td>
<td>11.00</td>
<td>2.75</td>
</tr>
<tr>
<td>pH</td>
<td>7.31</td>
<td>7.73</td>
<td>8.48</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Figure 4. Interpolation maps of (a); pH and (b); EC dS m⁻¹; and (c) map of soil texture.
3.2. Qualitative Description of the Soil Spectra

Spectral reflectance characteristics (Figure 5) aid in the identification of the main parameters of the soil, including organic matter, clay content, moisture content, cation exchange capacity, soil color, effect of soil minerals, soil surface condition, salinity, and alkalinity. The spectra library was created using reflectance spectra produced by laboratory experiments, and reflectance transformations were applied to further analyses of the spectra. In the Vis-NIR range, all of the soil samples showed distinctive reflectance spectra that corresponded to different molecular groups or chemical and/or mineralogical components [64]. There were noticeable absorption peaks and regions at wavelengths of 1400, 1900, and 2200 nm. Lower reflectance values were seen in the visible region (400–700 nm), which increased gradually in the near-infrared region (700–2500 nm). The hydroxyl group of free water is linked to the peaks at 1400 and 1900 nm, while clay minerals and lattice–OH features are linked to the absorption features at 1400 and 2200 nm [14,44]. Furthermore, the 2200 nm characteristic is linked to different organic functional groups in different clay minerals, such as CH₂, CH₃, etc., SiOH bonds, or cation–hydroxyl bonds [65] (for various layers, the OH stretching absorption feature was measured between 1410 and 1415 nm, the H–O–H absorption feature between 1910 and 1920 nm, and the Al–OH stretching between 2200 and 2210 nm). Furthermore, the absorption characteristics linked to the soil organic matter components were detected between 2390 and 2400 nm [66].

![Figure 5. Reflectance spectra of the examined soil samples.](image)

The average spectral library of CEC values is used for the different reflectance conversions. Figure 6 provides examples of various modifications. In addition to their impacts on strengthening the various absorption properties, the various transformations also served to lessen the various kinds of sounds in spectra.

![Figure 6. Cont.](image)
3.3. Prediction of Soil Properties Using PLSR

The most widely used multivariate quantifiable technique in soil research for chemometrics is PLSR, according to [67]. In order to assess soil attributes, this strategy adjusts soil reflectance [44]. It is superior to traditional methods for handling multi-collinearity in high-dimensional data. From each partial least square analysis, the following values were obtained: regression coefficients ($R^2$), residual prediction deviation (RPD), root mean square error (RMSE), and the ratio of performance to inter-quartile range (RPIQ). The coefficient of determination, residual predictive deviation, root mean square error, and performance-to-IQ ratio are commonly used assessment metrics for Vis-NIR model analysis [68–70]. $R^2$ and RMSE are useful metrics for assessing the models’ accuracy. Since these models exhibit similar stability at the same RPD level, RPD and RPIQ can be used to analyze the stability [46].

The PLSR model was also used to estimate the cross-validation segments, correlation loadings factor, and scatter plots of the measured vs. predicted soil parameters (Table 4 and Figure 7). Although they are less common [17], predictions of secondary attributes using Vis-NIR data have also been reported in a number of other temperate soil studies, including those by Abdul Munnaf [16] for pH, ex-Na, ex-Ca, ex-Mg, and ex-K ($0.58 \leq R^2 \leq 0.81$), Mouazen and Kuang [71] for ex P ($R^2 = 0.60$), and Ulusoy et al. [2] for CEC ($0.70 \leq R^2 \leq 0.75$).

Table 4. CEC prediction models’ performance with the matching preprocessing transformation.

<table>
<thead>
<tr>
<th>No.</th>
<th>Spectral Transformation</th>
<th>Calibrated (N = 73)</th>
<th>Validated (N = 31)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>RMSE</td>
<td>RPD</td>
</tr>
<tr>
<td>1</td>
<td>1/Log R</td>
<td>0.98</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>R</td>
<td>0.95</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>1/$\sqrt{R}$</td>
<td>0.95</td>
<td>0.62</td>
</tr>
<tr>
<td>4</td>
<td>(1/Log R)$'$</td>
<td>0.94</td>
<td>0.65</td>
</tr>
<tr>
<td>5</td>
<td>Log R</td>
<td>0.94</td>
<td>0.69</td>
</tr>
<tr>
<td>6</td>
<td>$\sqrt{R}$</td>
<td>0.93</td>
<td>0.76</td>
</tr>
<tr>
<td>7</td>
<td>$R'$</td>
<td>0.92</td>
<td>0.77</td>
</tr>
<tr>
<td>8</td>
<td>1/$R'$</td>
<td>0.90</td>
<td>0.85</td>
</tr>
<tr>
<td>9</td>
<td>($\sqrt{R}$)$'$</td>
<td>0.91</td>
<td>0.82</td>
</tr>
<tr>
<td>10</td>
<td>(Log R)$'$</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>11</td>
<td>(1/$\sqrt{R}$)$'$</td>
<td>0.82</td>
<td>1.16</td>
</tr>
<tr>
<td>12</td>
<td>(1/$R'$)</td>
<td>0.75</td>
<td>1.37</td>
</tr>
<tr>
<td>13</td>
<td>(1/Log R)$''$</td>
<td>0.75</td>
<td>1.37</td>
</tr>
</tbody>
</table>

Figure 6. Examples of various reflectance transformations used in the study.
Table 4. Cont.

<table>
<thead>
<tr>
<th>No.</th>
<th>Spectral Transformation</th>
<th>Calibrated (N = 73)</th>
<th>Validated (N = 31)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$R^2$</td>
<td>RMSE</td>
</tr>
<tr>
<td>14</td>
<td>$R''$</td>
<td>0.58</td>
<td>1.78</td>
</tr>
<tr>
<td>15</td>
<td>$(\sqrt{R})''$</td>
<td>0.53</td>
<td>1.88</td>
</tr>
<tr>
<td>16</td>
<td>$(\text{Log } R)''$</td>
<td>0.52</td>
<td>1.90</td>
</tr>
<tr>
<td>17</td>
<td>$(1/\sqrt{R})''$</td>
<td>0.28</td>
<td>2.27</td>
</tr>
<tr>
<td>18</td>
<td>$(1/R)''$</td>
<td>0.42</td>
<td>5.25</td>
</tr>
</tbody>
</table>

Figure 7. The best-performing models for CEC based on different transformation methods.
Table 4 provides a statistical summary of the PLSR models created for CEC prediction using different spectral modifications. Only those modifications that, after calibration and cross-validation, produced models with a moderate-to-high coefficient of determination ($R^2 > 0.6$) are included here. These were chosen based on the cross-validation procedure’s outcomes; namely, the lowest RMSE values and highest $R^2$, RPD, and RPIQ values. Out of the 18 reflectance transformations that were combined with reflectance spectra, 13 produced results that were adequate for both the model calibration and validation stages (Table 4). Conversely, the models that were produced by the second derivative spectrum of R, Log R, $\sqrt{R}$, reciprocal of R, and root R were not well calibrated. Based on the values of $R^2$, RMSE, RPD, and RPIQ in the calibration and validation datasets, the best fits of the 13 reflectance transformations were chosen. Ten transformation techniques produced calibration models for the current study that were satisfactory (Table 4), with $R^2 > 0.6$, RPD > 2.5, and RPIQP > 4.05. The reciprocal of Log R transformation ($R^2 = 0.98$, RMSE = 0.40, RPD = 6.99, and RPIQ = 9.22) was the most effective transformation and calibration model for assessing CEC in Vis-NIR spectra. This suggests that the best model for predicting CEC values is produced by combining the reciprocal of Log R with PLSR.

The RPD values of three transformations, namely, $(1/R)'$, $(1/\sqrt{R})'$, and $(1/\log R)'$, ranged between 2 and 2.5, indicating that employing these transformations resulted in excellent modeling of CEC using PLSR analysis. In contrast, the performance of the second derivative of R, Log R, root R, and reciprocal of root R was mediocre. The modeling is really bad because only the second derivative of the $(1/R)$ transformation has an RPD smaller than one.

3.4. CEC Mapping Based on Semivariogram Analysis

The degree to which the results are acceptable and compatible with previous studies is demonstrated by the accuracy of the spatial distribution map of CEC created using OK [72]. The semivariogram function in the context of the geostatistical analysis examines the null hypothesis, according to which the soil variable does not show spatial dependency at the selected lag h. The spatial dependence for CEC in the research area (Table 5) and the small nugget semivariance indicate that the lag h appears to characterize the spatial variation, and the lag h scale well matched the spatial variation of CEC. Therefore, it is not necessary to sample this variable again in order to identify spatial dependence.

<table>
<thead>
<tr>
<th>Spatial Distribution Model</th>
<th>$R^2$</th>
<th>RMSE</th>
<th>MSE</th>
<th>RMSSE</th>
<th>Range (m)</th>
<th>Nugget</th>
<th>Partial Sill</th>
<th>Sill</th>
<th>Nugget/Sill Ratios</th>
<th>Spatial Dependence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spherical</td>
<td>0.96</td>
<td>2.55</td>
<td>-0.003</td>
<td>1.018</td>
<td>14,846.4</td>
<td>4.78</td>
<td>2.35</td>
<td>7.13</td>
<td>0.67</td>
<td>Moderate</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.95</td>
<td>2.56</td>
<td>-0.005</td>
<td>1.025</td>
<td>13,878.2</td>
<td>3.68</td>
<td>3.53</td>
<td>7.21</td>
<td>0.51</td>
<td>Moderate</td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.88</td>
<td>2.55</td>
<td>-0.004</td>
<td>1.018</td>
<td>12,886.0</td>
<td>2.21</td>
<td>1.93</td>
<td>4.14</td>
<td>0.53</td>
<td>Moderate</td>
</tr>
<tr>
<td>Circular</td>
<td>0.94</td>
<td>2.55</td>
<td>-0.004</td>
<td>1.017</td>
<td>13,785.0</td>
<td>4.91</td>
<td>2.24</td>
<td>7.14</td>
<td>0.68</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

In situations when there is a high or moderate spatial correlation between the distributions of soil attributes, the range of the semivariogram indicates the average extent of these patches. Table 5 indicates that for the models—spherical, exponential, Gaussian, and circular models—the RMSSE values were 1.018, 1.025, 1.018, and 1.019, and the MSE values were $-0.003, -0.005, -0.004$, and $-0.004$, accordingly. For all parameters, the RMSSE values were near one and the MSE values were near zero [73]. This suggests that OK was a suitable and trustworthy method for estimating the CEC’s spatial distribution in the research area [72].

At the large scale, CEC values displayed a moderate geographic correlation, indicating that both extrinsic and intrinsic influences were present. Because of their strong spatial dependence, it was therefore reasonable to forecast the spatial distribution of CEC at a wide scale. Furthermore, there were minor variations within the ranges of different models of spatial distribution. Extrinsic (soil management techniques, such as fertilization) and
intrinsic (soil formation factors, like soil parent materials) factors can both have an impact on the spatial variability of soil parameters. According to [53,54], intrinsic factors are typically responsible for soil qualities that exhibit great geographical dependency, while extrinsic factors are responsible for those that exhibit weak spatial dependency. A larger range indicates that observed values of the CEC variable are influenced by other values of this variable over greater distances than soil variables, which have smaller ranges [74]. This shows that, at longer distances, CEC levels had an impact on nearby sand values. In the study region, Figure 8 shows the regional distribution of the CEC.

![Figure 8. The spatial distribution of the CEC throughout the study region.](image)

### 4. Conclusions

Comprehending the principles of cation exchange facilitates well-informed decision-making about soil management, fertilization with nutrients, preservation of the total fertility of agricultural and natural ecosystems, and attaining agricultural sustainability. Vis-NIR spectroscopy is an extremely useful method for CEC evaluation. The modeling of the correlation between the spectral response and the CEC concentrations served as the foundation for the current investigation. To be more specific, the PLSR was used to map and forecast CEC in situ analysis and Vis-NIR spectroscopy using the reflectance spectra obtained in the dry conditions of thirty-one soil samples were chosen in order to confirm the accuracy of the findings. The spatial variety and features of the samples that were chosen at random comprised the study region. Ten transformation approaches produced calibration models, with $R^2 > 0.6$, $RPD > 2.5$, and $RPIQ > 4.05$. These results show that the methods met the study’s requirements. The reciprocal of Log R transformation ($R^2 = 0.98$, $RMSE = 0.40$, $RPD = 6.99$, and $RPIQ = 9.22$) was the most effective transformation and calibration model for assessing CEC in Vis-NIR spectra. This suggests that PLSR and the reciprocal of Log R can be combined to provide the best model for predicting CEC values. Four models—Gaussian, exponential, circular, and spherical—fit the CEC values the best. The techniques used here allow for the rapid and widespread use of a “quick”, inexpensive tool. We put forth the overall proposal that the efficient monitoring of soil conditions through the use of earth observation instruments can be further enhanced in the future on a multi-scale and multi-parameter scale by combining information derived from earth observation with socioeconomic factors such as population, economy, etc. This will provide...
an integrated system that can be operationally adjusted to support and enhance local government efficiency. Additional investigation is required to examine various datasets obtained from an increased number of fields with varying soil types, textural attributes, land uses, and climate conditions.


**Funding:** This research received no external funding.

**Data Availability Statement:** The data presented in this study are available in the manuscript.

**Conflicts of Interest:** The authors declare no conflicts of interest.

**References**

8. Takele, C.; Iticha, B. Use of infrared spectroscopy and geospatial techniques for measurement and spatial prediction of soil properties. *Heliyon* 2020, 6, e05269. [CrossRef] [PubMed]
18. Nocita, M.; Stevens, A.; Noon, C.; van Wesemael, B. Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy. *Geoderma* 2013, 199, 37–42. [CrossRef]


22. Conforti, M.; Castrignano, A.; Robustelli, G.; Scariglia, F.; Stelluti, M.; Buttafuco, G. Laboratory-based Vis–NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content. *CATENA* 2015, 124, 60–67. [CrossRef]


27. Kaya, F.; Mishra, G.; Francaviglia, R.; Keshavarzi, A. Combining Digital Covariates and Machine Learning Models to Predict the Spatial Variation of Soil Cation Exchange Capacity. *Land* 2023, 12, 819. [CrossRef]


44. Viscarra Rossel, R.A.; Walvoort, D.J.J.; McBratney, A.B.; Janik, L.J.; Skjemstad, J.O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* 2006, 131, 59–75. [CrossRef]


52. Abuzaid, A.S.; Jhin, H.S.; Shokr, M.S.; El Baroudy, A.A.; Mohamed, E.S.; Rebouh, N.Y.; Bassouny, M.A. A Novel Regional-Scale Assessment of Soil Metal Pollution in Arid Agroecosystems. Agronomy 2023, 13, 161. [CrossRef]


57. Adeline, K.; Gomez, C.; Gorretta, N.; Roger, J.-M. Predictive ability of soil properties to spectral degradation from laboratory Vis-NIR spectroscopy data. Geoderma 2016, 218, 153. [CrossRef]


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.