forests

MDPI

Article

Cross-Site Prediction of Soil Organic Carbon in Plantation Forests Using Vis-NIR Spectroscopy and Target-Spectrum-Guided Source-Domain Weighting

Yun Deng [1,2] and Zubo Meng [1,2,3,*]

[1] Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, 12 Jiangan Road, Guilin 541004, China; 2002078@glut.edu.cn
[2] College of Computer Science and Engineering, Guilin University of Technology, 319 Yanshan Street, Yanshan District, Guilin 541006, China
[3] Guangxi Academy of Artificial Intelligence, 9 Jinliang Road, Liangqing District, Nanning 530200, China
* Correspondence: 2120241267@glut.edu.cn

Abstract

Soil organic carbon (SOC) is an important indicator of plantation-forest soil quality, but Vis-NIR models calibrated at one site may show systematic bias at another. This study used 370 Gaofeng samples as the source domain and 119 Yachang samples as the target domain to evaluate SOC-gradient-constrained spectral similarity weighting (SOC-SSW). Unlabeled target spectra guided source-sample weighting, while source SOC strata maintained calibration-gradient coverage. In Gaofeng-to-Yachang transfer, direct PLSR produced R^2 = 0.809, RMSE = 6.676 g kg^{-1}, and bias = −3.842 g kg^{-1}. SOC-SSW increased R^2 to 0.855, reduced RMSE to 5.800 g kg^{-1}, and decreased absolute bias by 75.42%. Paired bootstrap analysis confirmed lower RMSE and MAE, and paired absolute errors remained significantly lower after Holm correction (p = 0.0018). Similar accuracy was retained when weights were constructed from 10 or 40 target spectra, although the complete target set yielded the smallest absolute bias. Reverse transfer was unsuccessful, indicating direction-dependent applicability. SOC-SSW can therefore support conditional reuse of laboratory soil spectral datasets when target-site SOC labels are unavailable, provided that the source dataset is sufficiently representative.

Keywords: soil organic carbon; plantation forest; Vis-NIR spectroscopy; cross-site prediction; instance weighting; SOC-SSW; PLSR

Academic Editors: Aleksandar Dj Valjarević and Vincenzo Giannico

Received: 7 July 2026
Revised: 3 August 2026
Accepted: 11 August 2026
Published: 12 August 2026

1. Introduction

Soil organic carbon (SOC) is a key indicator for evaluating forest soil quality, nutrient cycling, and ecosystem carbon sequestration [1,2]. Plantation forests support timber production, ecological restoration, and regional carbon sinks; rapid SOC assessment is therefore important for soil-quality evaluation and forest management. Conventional SOC determination relies on field sampling and laboratory chemical analysis, which are time-consuming and costly and cannot easily meet the requirements of multi-site, large-scale, and high-frequency monitoring [3,4]. Hence, rapid, low-cost, and non-destructive SOC prediction is of practical importance [5].

Visible-near-infrared (Vis-NIR) spectroscopy continuously records soil reflectance information and indirectly reflects SOC variations through the integrated responses of soil color, organic matter, mineral composition, moisture status, and particle-size composition.

It has been widely used to predict SOC and other soil properties [6–10]. In soil spectral modeling, PLSR, random forest (RF), support vector regression (SVR), and deep learning methods have often been used to characterize relationships between spectral features and soil properties, including SOC, soil organic matter, and soil water content [11,12]. However, the transferability of existing models across different forest farms remains an important limitation to the broader application of soil spectroscopy.

The spectral response of SOC in plantation forests is influenced not only by organic carbon content, but also by soil texture, mineral composition, soil color, stand structure, management history, and pedogenic environment. When source and target forest farms differ in soil background, SOC distribution, or spectral response, a model developed from the source forest farm may show reduced accuracy or systematic bias when directly applied to the target forest farm. Such bias can affect the reliability of SOC-level assessment and soil carbon monitoring in the target forest farm. To address cross-regional or cross-site prediction problems, previous studies have used target-domain calibration samples, spectral-library localization, spiking, transfer learning, or model updating to improve model generalization [13–17]. Recent reviews and localization studies have further shown that the spectral similarity between calibration samples and the target domain, as well as the SOC coverage of the training set, strongly influences cross-regional prediction performance in soil spectroscopy [18–20]. However, the use of measured SOC samples from the target domain increases sampling and laboratory costs, which conflicts with the goal of rapid and low-cost monitoring. These approaches differ in their use of target-domain information. Spiking and most local-calibration strategies incorporate target samples with measured SOC, whereas unsupervised domain adaptation and instance weighting use unlabeled target spectra to adjust source features or source-sample contributions. Calibration transfer mainly addresses instrument- or protocol-induced spectral differences, while the present study focuses on site-related distribution differences under a unified laboratory measurement workflow.

In practical forest soil carbon monitoring, a new target forest farm may have only Vis-NIR spectra but no SOC or other physicochemical measurements. Therefore, improving the transferability of a source-domain model to the target forest farm using only the unlabeled spectral distribution of the target domain, without using target-domain SOC labels, is a key issue for the cross-site reuse of existing soil spectral databases. Source-domain sample selection or weighting guided by the target-domain spectral distribution is a direct and feasible strategy [16,18–20]. Further introducing a source-domain SOC-gradient constraint can help avoid excessive concentration of training samples within a narrow SOC range and thereby improve the stability of transfer prediction. Accordingly, SOC-SSW should be regarded as an instance-weighting strategy that combines target-spectrum similarity with source-domain SOC-gradient balancing, rather than as a new regression algorithm.

This study used plantation-forest soil samples from Gaofeng and Yachang Forest Farms in Guangxi, China, as the source and target domains, respectively. Without using Yachang SOC labels for model training, spectral-similarity calculation, or sample weighting, we compared direct source-domain transfer, target-domain spectral similarity selection, PCA-CORAL distribution alignment, and SOC-SSW. The objectives were to: (1) evaluate the accuracy and systematic bias of direct cross-site transfer; (2) analyze the role of unlabeled target spectra in constraining source-domain sample selection and weighting; and (3) assess whether SOC-SSW improves target-site SOC prediction without target-domain physico-chemical data. The results provide methodological support for rapid SOC monitoring and cross-site reuse of existing laboratory soil spectral databases.

2. Materials and Methods

To support rapid SOC monitoring in plantation forests, this study developed a source-domain-assisted Vis-NIR prediction workflow comprising soil sampling, laboratory spectral acquisition, unified preprocessing, source-target partitioning, cross-site scenario design, model training, and target-site evaluation. Gaofeng was used as the source-domain dataset and Yachang as the target-domain dataset to test whether an existing laboratory soil spectral dataset could support prediction under an unlabeled-target setting.

2.1. Study Area and Data Collection

This study was conducted in two subtropical plantation forest areas in southern China, namely Gaofeng Forest Farm and Yachang Forest Farm in Guangxi (Figure 1). The cross-site SOC prediction experiment was based on laboratory Vis-NIR proximal spectral data. The study area has favorable hydrothermal conditions, diverse vegetation types, and concurrent soil organic matter accumulation and mineral weathering processes, resulting in spatial heterogeneity. Random sampling was adopted to select sampling sites within each forest farm, and surface soil samples (0–20 cm depth) were collected, yielding 370 valid samples from Gaofeng Forest Farm and 119 from Yachang Forest Farm.

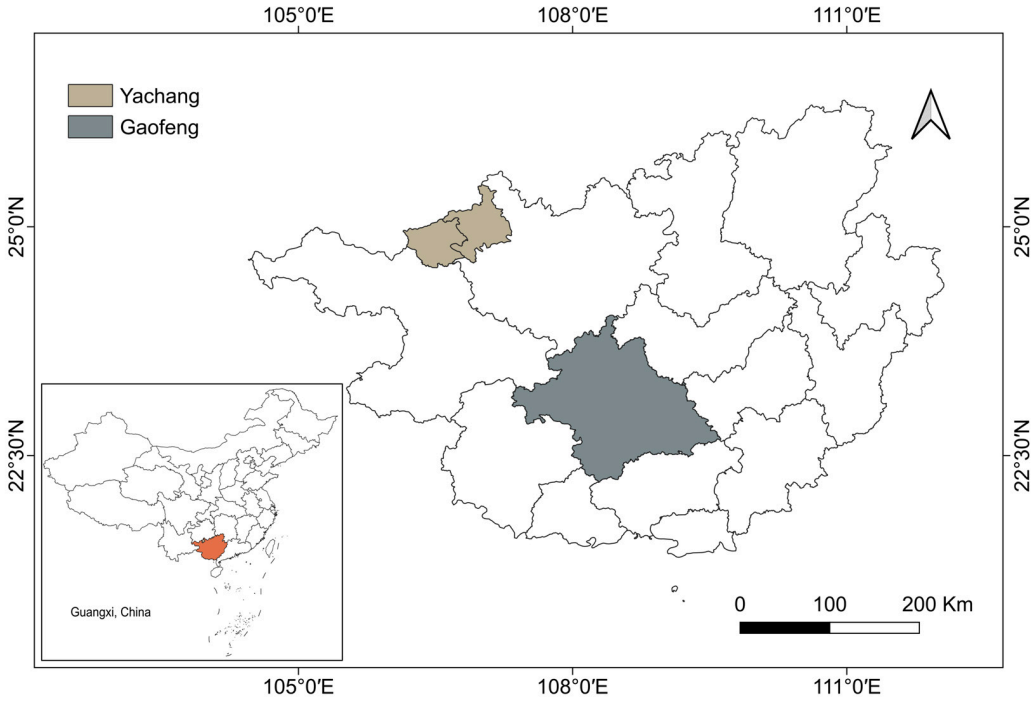

Figure 1. Location of the study area and the relative positions of Gaofeng and Yachang Forest Farms in Guangxi, China.

After being transported to the laboratory, soil samples were air-dried, plant residues and gravel were removed, and the samples were ground and passed through a 2 mm sieve to reduce the effects of particle-size differences and sample heterogeneity on spectral measurement. A portion of each processed sample was used for SOC determination, while another portion was used for laboratory Vis-NIR spectral acquisition (Figure 2). SOC content was determined using the potassium dichromate oxidation method. Soil reflectance spectra were collected using an ASD FieldSpec 4 Hi-Res portable spectroradiometer (Analytical Spectral Devices Inc., Boulder, CO, USA) over the wavelength range of 350–2500 nm, and reflectance data were exported at 1 nm intervals. Each sample was scanned five times at different positions, and the average spectrum was used as the original reflectance spectrum of that sample. Recent forest SOC mapping and soil spectroscopy studies em-

phasize the need for consistent spectral data acquisition and preprocessing workflows to ensure comparability [21]. In this study, the larger Gaofeng dataset was used as the source domain and the Yachang dataset as the target domain. The spectra and SOC data from Gaofeng were used for model training, source-domain internal validation, and transfer modeling; the Yachang samples provided only their unlabeled spectral distribution during transfer modeling, and their SOC measurements were used only for the final independent evaluation of prediction performance.

Figure 2. Laboratory Vis-NIR spectral acquisition of soil samples.

2.2. Vis-NIR Spectral Preprocessing

All spectra were preprocessed using a unified combination of Savitzky–Golay smoothing and standard normal variate transformation, hereafter referred to as SG-SNV. The SG smoothing window length was 11, the polynomial order was 2, and the derivative order was 0; similar smoothing and scatter-correction procedures are widely used in recent Vis-NIR SOC studies [22,23]. Each spectrum was then transformed using SNV to reduce scattering effects and baseline shifts caused by particle size, sample surface condition, and optical-path differences. Neither SG nor SNV used SOC label information, and SNV was applied independently to each spectrum without using any test-set statistics. The same preprocessing workflow was used in all modeling scenarios to ensure that differences among scenarios were not caused by different preprocessing methods. The comparison between the raw and SG-SNV mean spectra was used as a descriptive visualization only. No formal domain-discrepancy statistic was used to claim a quantitative reduction in source-target difference. The numbers of PCA components and nearest neighbors used in the subsequent similarity calculation were fixed before examining target SOC performance. All wavelengths from 350 to 2500 nm were retained for modeling, and no wavelength intervals were removed separately. The SG-SNV configuration and the use of the complete wavelength range were predefined before transfer evaluation and were applied uniformly to both domains. They were not selected using target-domain SOC performance, and no claim is made that this preprocessing configuration is superior to alternative derivative, detrending, or scatter-correction methods.

2.3. Cross-Site SOC Prediction Modeling Scenarios

2.3.1. Scenario A: Source-Domain Internal Validation

Scenario A was designed as source-domain internal validation. The Gaofeng source-domain samples were split into approximately 50% training and 50% testing subsets, and the procedure was repeated 20 times to evaluate the modeling feasibility and stability of the relationship between Gaofeng SOC and Vis-NIR spectra. This scenario did not involve the Yachang target domain and was used only as a reference for source-domain data usability and model stability.

2.3.2. Scenario B: Direct Source-Domain Transfer

Scenario B represented direct source-domain transfer. All Gaofeng samples were used to train the model, which was then applied directly to all Yachang samples. Yachang SOC values were used only to calculate prediction metrics and were not involved in model training. This scenario assessed direct transferability and systematic bias when no target-domain physicochemical data were available.

2.3.3. Scenario C: Target-Domain Spectral-Guided Source-Domain Similarity Selection Transfer

Scenario C used only the unlabeled spectral distribution of the target domain to select source-domain samples that were spectrally closer to the target forest farm. Let X_s denote the preprocessed spectral matrix of the Gaofeng source domain, y_s the measured SOC values of the source domain, and X_t the preprocessed unlabeled spectral matrix of the Yachang target domain. First, X_s and X_t were combined and standardized. Principal component analysis (PCA) was then performed on the combined spectral matrix [24], and the first 20 principal components were extracted to obtain representations Z_s and Z_t of the source and target domains in a common latent-variable space.

In the latent-variable space, a nearest-neighbor model was built using the target-domain samples, and Euclidean distance was used to calculate the distance from each source-domain sample to the target-domain spectral distribution. For the i-th source-domain sample, its five nearest target-domain neighbors were selected, and the mean distance to these five neighbors was defined as its similarity distance:

$$d_i = \frac{1}{K}\sum_{j=1}^{K}\left\| z_i^s - z_{t,NN_j(i)} \right\|_2 \quad (1)$$

where z_i^s is the feature vector of the i-th source-domain sample in PCA space and $z_{t,NN_j(i)}$ is its j-th nearest target-domain sample. A smaller d_i indicates that the source-domain sample is closer to the unlabeled target-domain spectral distribution.

Source-domain samples were ranked by increasing similarity distance, and predefined retention ratios of 40%, 50%, 60%, and 70% were evaluated. These ratios were treated as an exploratory sensitivity analysis rather than as target-label-free optimized hyperparameters. Target-domain SOC values were not used to select an operational retention ratio. Consequently, the 40% result is reported only as the best observed similarity-selection result for the present Gaofeng-to-Yachang combination, while all four ratios are shown to demonstrate the dependence of sample-selection performance on the retained proportion.

2.3.4. Scenario D: PCA-CORAL Spectral Distribution Alignment Transfer

Scenario D used PCA-CORAL for unsupervised spectral distribution alignment. CORAL reduces the discrepancy in second-order statistics between domains by aligning the covariance matrices of source and target features [25]. This scenario did not use target-domain SOC labels and used only the distributions of source and target-domain

spectral features. Because the original Vis-NIR spectra had high dimensionality, direct covariance alignment could lead to numerical instability. Therefore, source and target-domain spectra were first combined, standardized, and reduced to the first 50 principal components to obtain feature matrices Z_s and Z_t in a common PCA space.

Let μ_s and μ_t be the mean vectors of Z_s and Z_t, respectively, and C_s and C_t their covariance matrices. To improve numerical stability in matrix decomposition, a regularization term was added to each covariance matrix:

$$C_s = \mathrm{cov}(Z_s - \mu_s) + \epsilon I \tag{2}$$

$$C_t = \mathrm{cov}(Z_t - \mu_t) + \epsilon I \tag{3}$$

where $\epsilon = 10^{-6}$ and I is the identity matrix. Eigenvalue decomposition was then used to calculate $C_s^{-1/2}$ and $C_t^{1/2}$, and the source-domain PCA features were aligned to the covariance structure of the target domain as follows:

$$Z_s^{coral} = (Z_s - \mu_s)C_s^{-1/2}C_t^{1/2} + \mu_t \tag{4}$$

After alignment, the transformed source-domain features Z_s^{coral} and their measured SOC values y_s were used to train the regression model, and the target-domain PCA features Z_t were used for SOC prediction. This scenario was used to test the effect of unsupervised spectral covariance alignment on cross-site SOC prediction. Except for Scenario D, all other scenarios used the SG-SNV-preprocessed Vis-NIR spectra as regression features.

2.3.5. Scenario E: SOC-Gradient-Constrained Spectral Similarity Weighting Transfer

Scenario E was the proposed SOC-gradient-constrained spectral similarity weighting method (SOC-SSW), and its overall workflow is shown in Figure 3. This method used the same source-target spectral similarity distance as Scenario C, but instead of directly discarding source-domain samples, it constructed sample weights by combining target-domain spectral similarity with the source-domain SOC-gradient distribution. Sample weighting can be regarded as a form of importance weighting for addressing distribution differences between the source and target domains [26,27]. Source-domain samples that were closer to the target-domain spectral distribution contributed more to model training, whereas SOC stratification prevented the training samples from being overly concentrated within a single SOC level and maintained coverage of the source-domain SOC gradient.

First, the similarity distance of each source-domain sample was converted into a Gaussian-kernel similarity weight:

$$s_i = exp\left(-\frac{d_i^2}{2\sigma^2}\right) \tag{5}$$

where d_i is the mean nearest-neighbor distance from the i-th source-domain sample to the unlabeled target-domain spectral distribution, and σ is the kernel bandwidth. In this study, the median of all non-zero d_i values was used as σ. To avoid numerical instability caused by extremely small weights, the lower bound of the similarity weight was set to 10^{-6}.

Second, the Gaofeng source-domain samples were stratified by their measured SOC values using equal-frequency stratification. Specifically, the source-domain samples were sorted by SOC from low to high and divided into five SOC-gradient strata based on quantiles. Target-domain SOC labels were not involved in this stratification process. Within each SOC stratum, the similarity weights were normalized and the total weight of each

SOC stratum was set to be equal. If the i-th source-domain sample belonged to the g-th SOC stratum, its stratified similarity weight was calculated as:

$$w_i' = \frac{s_i}{\sum_{j \in G_g} s_j} \times \frac{1}{G} \tag{6}$$

where G_g denotes the set of source-domain samples in the g-th SOC stratum and $G = 5$ is the number of SOC strata. This operation ensured that different SOC-gradient strata contributed equally to the overall training weights, while the relative contribution of samples within the same SOC stratum was determined by their spectral similarity to the target domain.

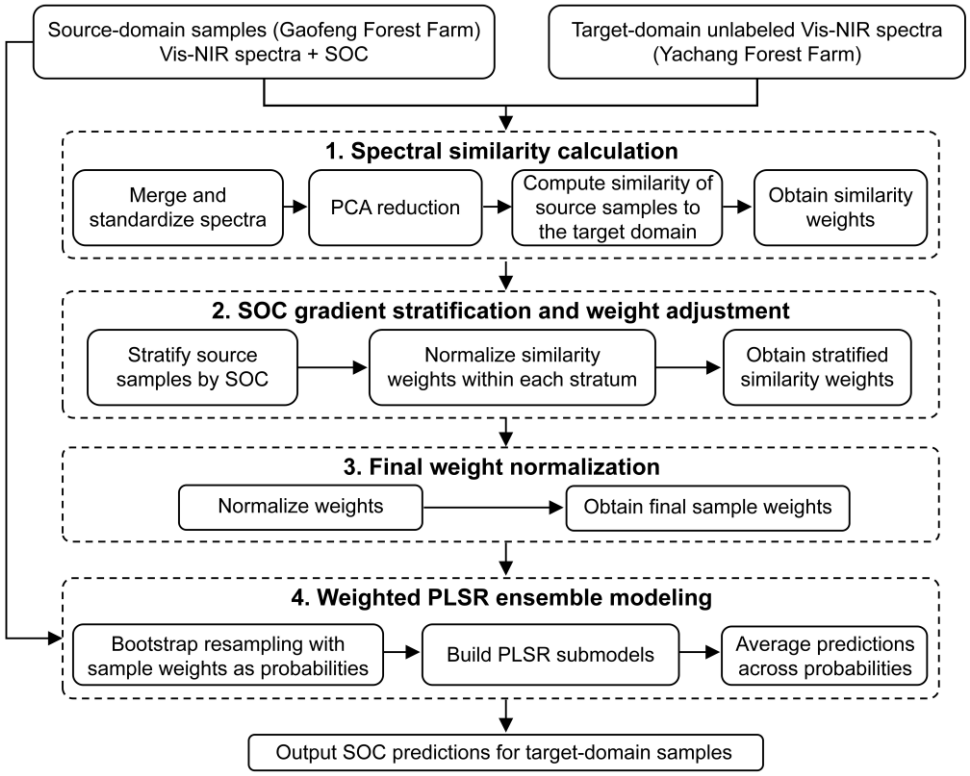

Figure 3. Workflow of SOC-SSW.

Third, the sample weights were further standardized to have a mean of 1:

$$w_i = \frac{w_i'}{\frac{1}{n_s}\sum_{i=1}^{n_s} w_i'} \tag{7}$$

where n_s is the number of source-domain samples and w_i is the final sample weight used for model training.

Because PLSR does not directly support sample weights, this study implemented PLSR sample weighting using a weighted bootstrap ensemble strategy. Specifically, the normalized sample weights were used as sampling probabilities to resample, with replacement, a training set of the same size as the source domain. This process was repeated to build 50 PLSR submodels. Bootstrap ensemble modeling can improve model robustness through repeated resampling and model averaging [28]. The number of latent variables in each PLSR submodel was determined by five-fold cross-validation within the corresponding training set, with an upper limit of 30 latent variables. The final prediction of Scenario E was obtained by averaging the predictions of the 50 submodels for the target-domain samples.

For RF and SVR, the sample weights described above were used directly because both models accept sample weights. In all scenarios, target-domain SOC values were used only for final performance evaluation and were not involved in similarity calculation, SOC stratification, sample weighting, or model training. The predefined Scenario C retention ratios were assessed only as an exploratory sensitivity analysis and were not treated as operationally optimized parameters.

2.3.6. Additional Robustness and Sensitivity Analyses

Two additional analyses were conducted to assess the robustness and practical applicability of the transfer framework. First, direction dependence was examined by reversing the source and target domains: the 119 Yachang samples were used as the labeled source domain and the 370 Gaofeng spectra as the unlabeled target domain. The same SG-SNV preprocessing, similarity-selection ratios, PCA-CORAL procedure, SOC-SSW weighting rule, and PLSR configuration were retained, and Gaofeng SOC labels were used only for final evaluation. Second, because the principal SOC-SSW experiment used all 119 Yachang spectra to characterize the target distribution, random subsets of 10 and 40 target spectra were also used to construct the similarity weights. Each partial-target setting was repeated three times, and the resulting models predicted all 119 target samples; the complete-target condition corresponded to the principal SOC-SSW experiment. Target SOC labels were not used to select the subsets or calculate the weights.

To avoid target-label-driven tuning, the transfer parameters were fixed before target evaluation and applied consistently in both directions: 20 PCA components and five nearest neighbors for spectral similarity, five equal-frequency source SOC strata for SOC-SSW, and 50 PCA components for CORAL. These a priori settings were intended to balance local target representation, numerical stability, and adequate sample numbers within each SOC stratum. A comprehensive multi-parameter optimization was not conducted because the objective was to evaluate the transfer framework rather than optimize it for a single target site.

2.4. Prediction Models

In this study, PLSR, RF, and SVR were selected as representative regression models for SOC prediction. PLSR was used as the main transfer-evaluation model because it is suitable for high-dimensional and multicollinear soil spectral data. RF and SVR were used as nonlinear ensemble and kernel-based baseline models, respectively, to compare model responses to the distribution discrepancy between the source and target domains [29–31]. The number of PLSR latent variables was determined by five-fold cross-validation within each training set, with an upper limit of 30 latent variables; the test set was not involved in component selection. The RF model used 500 decision trees, max_features = "sqrt", min_samples_leaf = 3, and max_depth = 10. The SVR model used a radial basis function (RBF) kernel, with C = 10, gamma = "scale", and ε = 0.1. As representative baseline nonlinear machine-learning models, RF and SVR used preset empirical parameters to avoid excessive model-specific tuning. All models used the same SG-SNV-preprocessed spectral features and evaluation metrics to ensure comparability across transfer strategies. The focus of this study was not to develop a new regression model, but to evaluate the effects of different source-domain sample selection and weighting strategies on target-domain SOC prediction performance and systematic bias.

2.5. Model Evaluation Metrics and Statistical Analysis

Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), ratio of performance to deviation (RPD), ratio of performance to interquartile distance (RPIQ), and bias. Higher R^2, RPD, and RPIQ values indicate better prediction performance, whereas lower RMSE and MAE

values indicate lower prediction error. Bias was used to assess systematic overestimation or underestimation. RPD and RPIQ are commonly used for comprehensive evaluation of soil spectral prediction models [32]. The metrics were calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (9)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (10)$$

$$RPD = \frac{SD_y}{RMSE} \quad (11)$$

$$RPIQ = \frac{IQR_y}{RMSE} \quad (12)$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \quad (13)$$

where y_i and $\hat{y}_i$ are the measured and predicted SOC values of the i-th sample, respectively; $\bar{y}$ is the mean measured SOC value of the test set; n is the number of test samples; SD_y is the standard deviation of measured SOC in the test set; IQR_y is the interquartile range of measured SOC in the test set; a positive bias indicates overall overestimation, whereas a negative bias indicates overall underestimation.

Statistical uncertainty in the principal Gaofeng-to-Yachang transfer was quantified using 5000 paired bootstrap resamples of the target samples. Percentile 95% confidence intervals were calculated for R^2, RMSE, MAE, bias, RPD, and RPIQ, and paired bootstrap confidence intervals were calculated for differences between SOC-SSW and the principal baselines. Paired absolute errors were compared using the Wilcoxon signed-rank test, with Holm adjustment for the three comparisons of SOC-SSW against direct transfer, 40% similarity selection, and CORAL. An adjusted p value below 0.05 was considered statistically significant. Prediction errors were also summarized by quartiles of measured target SOC and by whether target SOC values fell below, within, or above the source-domain SOC range. For the target-spectrum availability analysis, the 10- and 40-spectrum settings are reported as means ± standard deviations across three random subsets; the complete 119-spectrum condition corresponds to the single principal SOC-SSW experiment.

ChatGPT (GPT-5.6 Sol; OpenAI) was used to assist with drafting and revising Python 3.9 scripts for the additional analyses and organizing statistical outputs. All code, analyses, and numerical results were independently reviewed and verified by the authors.

3. Results

3.1. Source-Target Differences and Baseline Transfer

The Gaofeng source-domain samples and the Yachang target-domain samples differed in both Vis-NIR spectral characteristics and SOC distribution (Figure 4). Gaofeng SOC ranged from 6.665 to 64.535 g kg^{-1}, with a mean of 19.984 g kg^{-1} and an SD of 8.517 g kg^{-1}. Yachang SOC ranged from 4.260 to 80.040 g kg^{-1}, with a mean of 22.379 g kg^{-1} and an SD of 15.321 g kg^{-1} (Table 1). The Yachang dataset therefore covered a wider SOC range and had a larger IQR than the Gaofeng dataset. In the raw reflectance spectra, the Yachang samples generally showed higher reflectance than the Gaofeng samples, with more evident separation in the near-infrared and shortwave-infrared regions. After SG-SNV preprocessing, visual inspection showed that the mean spectral shapes became more similar; however, this descriptive comparison did not quantify a reduction in source-target

domain discrepancy. Local differences nevertheless remained around 1400 nm, 1900 nm, 2200 nm, and the end of the shortwave-infrared region. The Gaofeng SOC values were concentrated mainly in the low-to-medium range, whereas Yachang covered a wider range and included more high-SOC samples. Thus, both spectral and SOC distributions differed between the source and target domains.

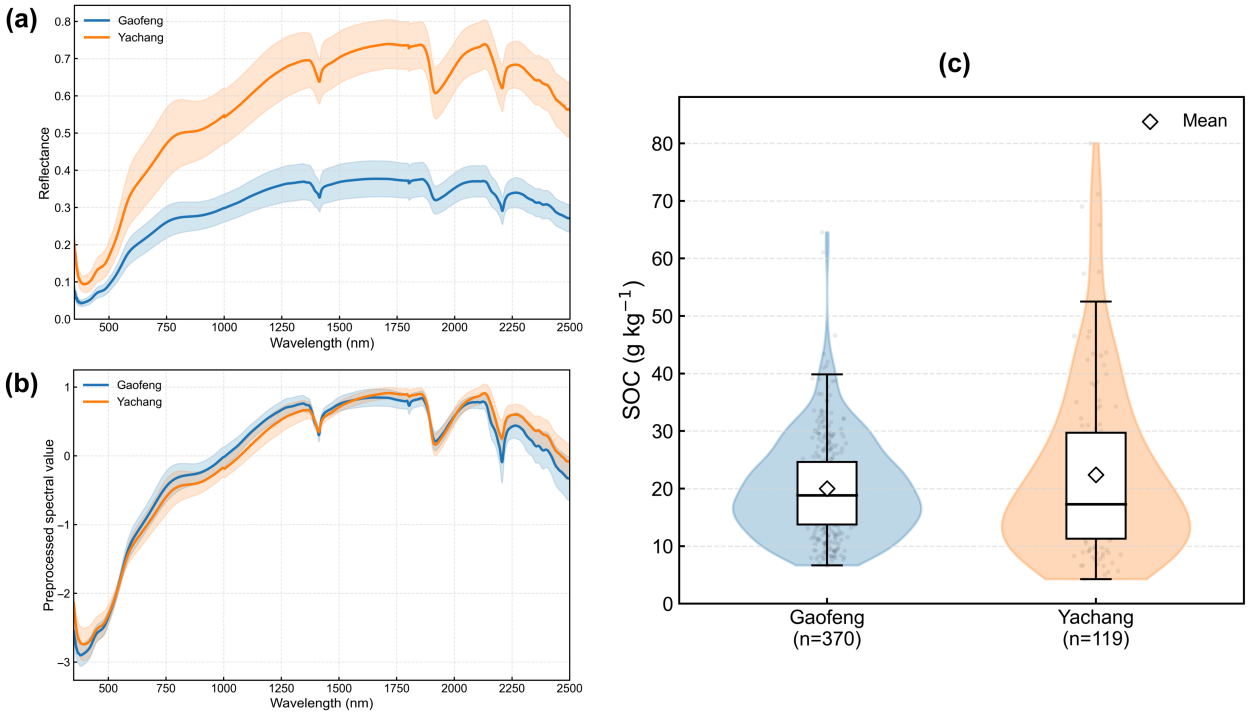

Figure 4. Spectral and SOC distribution characteristics of samples from Gaofeng and Yachang Forest Farms. (**a**) Raw Vis-NIR reflectance spectra; (**b**) SG-SNV-preprocessed Vis-NIR spectra; (**c**) SOC distribution. Solid lines indicate mean spectra, and shaded areas show variability among samples.

Table 1. Descriptive statistics of SOC in Gaofeng and Yachang Forest Farms.

Forest Farm	n	Minimum	Q1	Median	Mean	Q3	Maximum	SD	IQR
Gaofeng	370	6.665	13.757	18.827	19.984	24.623	64.535	8.517	10.866
Yachang	119	4.260	11.280	17.300	22.379	29.685	80.040	15.321	18.405

Note: SOC values are expressed in g kg^{-1}. Q1 and Q3 denote the first and third quartiles; IQR = Q3 − Q1.

The source-domain internal-validation results are presented in Table 2. PLSR showed the highest mean R^2 and the lowest RMSE and MAE, whereas RF and SVR produced similar results. Table 3 shows that model differences became more pronounced after transfer to Yachang. PLSR retained the best target-domain performance and captured the overall SOC trend, but its negative bias indicated systematic underestimation. RF and SVR produced substantially higher errors and weaker target-domain fits.

Table 2. Scenario A: Source-domain internal validation performance in Gaofeng Forest Farm.

Model	R^2	RMSE (g kg^{-1})	MAE (g kg^{-1})	RPD	RPIQ	Bias (g kg^{-1})
PLSR	0.772 ± 0.022	3.980 ± 0.192	3.019 ± 0.150	2.105 ± 0.099	2.729 ± 0.124	−0.366 ± 0.441
RF	0.648 ± 0.034	4.941 ± 0.242	3.523 ± 0.167	1.696 ± 0.083	2.199 ± 0.111	0.007 ± 0.323
SVR	0.645 ± 0.029	4.962 ± 0.208	3.351 ± 0.151	1.688 ± 0.068	2.188 ± 0.092	−0.662 ± 0.457

Note: Values are means ± standard deviations from 20 repeated splits. This table is used only to describe source-domain modeling feasibility and is not used to rank cross-site transfer scenarios.

Table 3. Scenario B: Prediction performance when directly transferring the source-domain model to Yachang Forest Farm.

Model	R^2	RMSE (g kg^{-1})	MAE (g kg^{-1})	RPD	RPIQ	Bias (g kg^{-1})
PLSR	0.809	6.676	5.165	2.295	2.757	−3.842
RF	0.382	11.991	9.084	1.278	1.535	3.167
SVR	0.371	12.097	7.564	1.266	1.521	−1.913

Note: Yachang SOC measurements were used only for final evaluation and were not involved in model training or parameter selection.

3.2. Source Sample Selection and PLSR Transfer Performance Under Target-Domain Spectral Constraints

The PCA score distribution based on SG-SNV-preprocessed spectra is shown in Figure 5. PC1 and PC2 explained 66.4% and 21.1% of the total spectral variance, respectively, accounting for 87.5% cumulatively. The Gaofeng source-domain samples and the Yachang target-domain samples partially overlapped in the principal component space, but their distributions were not completely consistent. Some Gaofeng samples were located near the distribution area of the Yachang samples, whereas others were distributed farther away from the target-domain samples. After similarity selection guided by unlabeled target-domain spectra, the retained source-domain samples were mainly concentrated in the overlapping region between the source and target domains. The final SOC-SSW weights were non-uniform. Within each SOC stratum, source-domain samples located closer to the target-domain spectral distribution generally received larger relative weights, while the normalization procedure maintained an equal total contribution from each SOC stratum. Figure 5b therefore provides a qualitative visualization of the concentration of highly weighted source samples in the source-target overlap region.

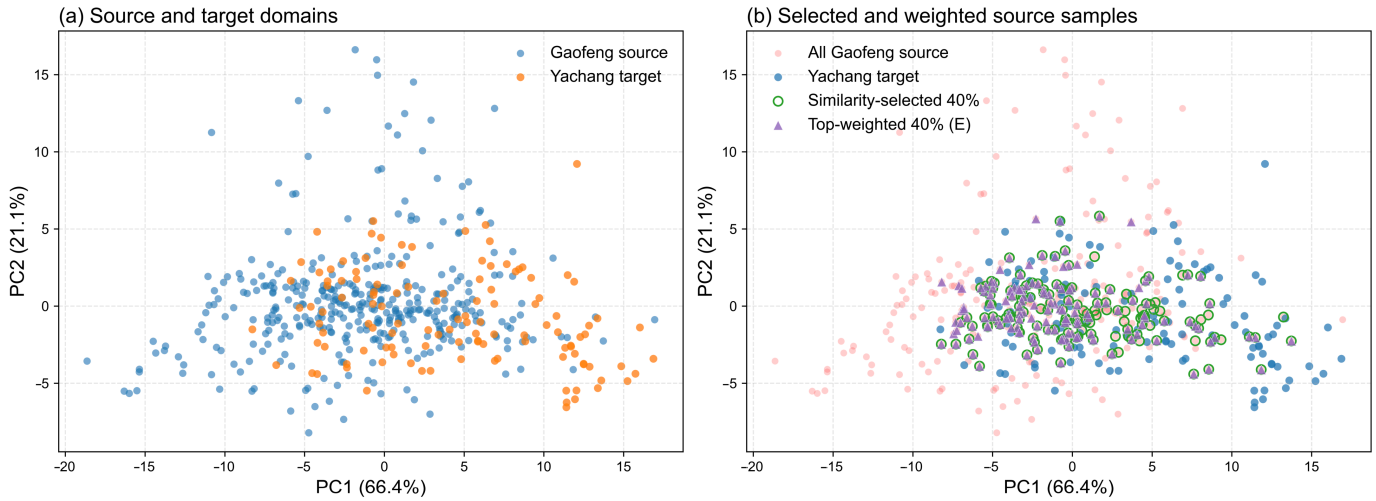

Figure 5. PCA score distribution based on SG-SNV-preprocessed Vis-NIR spectra. PC1 and PC2 explained 66.4% and 21.1% of the total spectral variance, respectively, accounting for 87.5% cumulatively. (**a**) Distribution of source-domain Gaofeng samples and target-domain Yachang samples; (**b**) similarity-selected samples and highly weighted source-domain samples guided by the target-domain spectra.

The PLSR prediction performance under different similarity selection ratios is shown in Table 4. The 40% retention setting had the lowest observed RMSE (6.627 g kg^{-1}) and the smallest absolute bias (1.719 g kg^{-1}) among the four ratios. The 50%, 60%, and 70% settings

produced RMSE values of 7.307, 9.272, and 8.513 g kg^{-1}, respectively. The 40% result was therefore retained in the main strategy comparison as the best observed exploratory similarity-selection baseline, not as an operational ratio selected without target labels. When the top 40% most similar source-domain samples were retained, PLSR achieved the best performance among the selection schemes, and the absolute Bias was markedly lower than that of direct transfer. The response to the retention ratio was non-monotonic: performance deteriorated at 50% and 60%, while the 70% setting recovered slightly but remained worse than the 40% setting. The target-domain prediction results therefore differed clearly among different source-sample retention ratios. The prediction results of different PLSR transfer strategies are shown in Table 5 and Figure 6. The direct transfer scheme captured the overall SOC variation trend in the Yachang target site, but the prediction points were generally lower than the 1:1 line. After 40% similarity selection, the underestimation was reduced. The PCA-CORAL distribution alignment scheme showed prediction points that deviated more clearly from the 1:1 line, with overall performance lower than the other PLSR schemes. The SOC-SSW scheme achieved the best overall performance across the evaluation metrics, and its prediction points were closer to the 1:1 line.

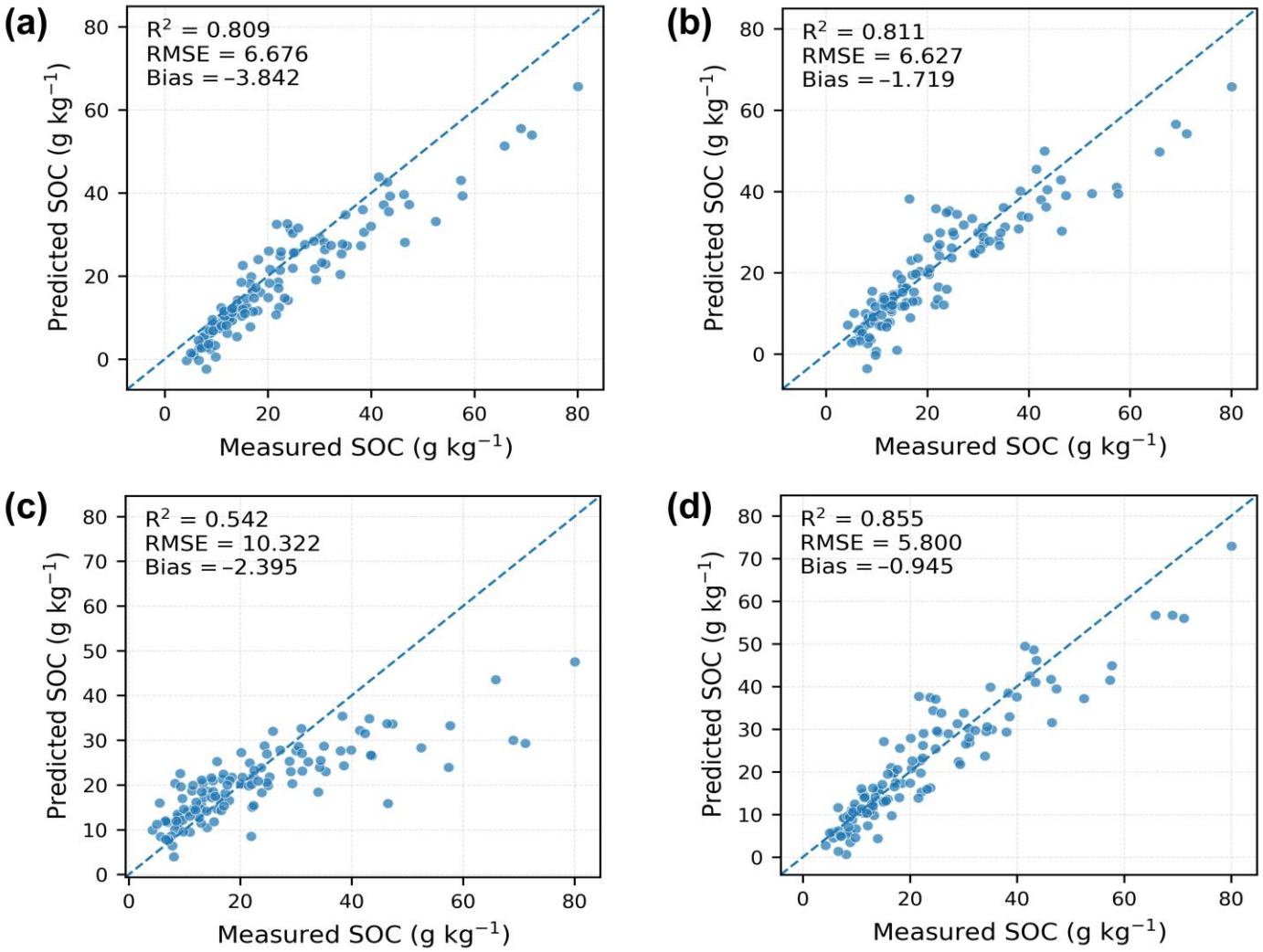

Figure 6. PLSR prediction results for SOC in Yachang Forest Farm under different transfer strategies. (**a**) Direct source-domain transfer; (**b**) similarity selection 40%; (**c**) CORAL distribution alignment; (**d**) SOC-SSW. The dashed line denotes the 1:1 fitting line.

Table 4. PLSR prediction performance under different source-domain similarity selection ratios in Scenario C.

Scenario C	R^2	RMSE (g kg^{-1})	MAE (g kg^{-1})	RPD	RPIQ	Bias (g kg^{-1})
Similarity selection 40%	0.811	6.627	4.910	2.312	2.777	−1.719
Similarity selection 50%	0.771	7.307	5.421	2.097	2.519	−3.495
Similarity selection 60%	0.631	9.272	7.439	1.652	1.985	−6.553
Similarity selection 70%	0.689	8.513	6.566	1.800	2.162	−5.555

Table 5. Comparison of PLSR prediction performance under different transfer strategies in the target forest farm.

Scenario	R^2	RMSE (g kg^{-1})	MAE (g kg^{-1})	RPD	RPIQ	Bias (g kg^{-1})
Scenario B: Direct source-domain transfer	0.809	6.676	5.165	2.295	2.757	−3.842
Scenario C: Similarity selection 40%	0.811	6.627	4.910	2.312	2.777	−1.719
Scenario D: CORAL distribution alignment	0.542	10.322	6.789	1.484	1.783	−2.395
Scenario E: SOC-SSW	0.855	5.800	4.309	2.642	3.173	−0.945

3.3. Responses of Different Regression Models to Transfer Strategies

The target-domain prediction performance of different regression models under the main transfer strategies is presented in Table 6. Overall, PLSR outperformed RF and SVR across all transfer schemes and maintained relatively stable predictive performance under direct transfer, similarity selection, PCA-CORAL distribution alignment, and SOC-SSW. RF and SVR showed generally higher prediction errors in the target domain, with relatively small differences between the two models. Under direct transfer, PLSR performed markedly better than RF and SVR in the target domain. After 40% similarity selection, RF showed a slight improvement compared with direct transfer, whereas SVR changed only slightly.

Table 6. Target-forest prediction performance of different regression models under the main transfer strategies.

Transfer Strategy	Model	R^2	RMSE (g kg^{-1})	MAE (g kg^{-1})	Bias (g kg^{-1})
Scenario B: Direct source-domain transfer	PLSR	0.809	6.676	5.165	−3.842
	RF	0.382	11.991	9.084	3.167
	SVR	0.371	12.097	7.564	−1.913
Scenario C: Similarity selection 40%	PLSR	0.811	6.627	4.910	−1.719
	RF	0.425	11.573	8.769	3.431
	SVR	0.308	12.688	8.140	−1.952
Scenario D: CORAL distribution alignment	PLSR	0.542	10.322	6.789	−2.395
	RF	0.362	12.182	8.428	−1.974
	SVR	0.377	12.040	7.847	−2.532
Scenario E: SOC-SSW	PLSR	0.855	5.800	4.309	−0.945
	RF	0.366	12.149	9.177	3.334
	SVR	0.374	12.074	7.208	−2.553

Note: Scenario A was source-domain internal validation and is not included in this table.

Under the PCA-CORAL distribution alignment scheme, none of the three models achieved better performance than their respective best transfer schemes, and the performance decline was more evident for PLSR. Under the SOC-SSW scheme, PLSR achieved the best prediction performance among all model-strategy combinations, whereas RF and

SVR did not show simultaneous improvement. The responses of different models to the transfer strategies varied, with PLSR showing the most evident response to SOC-SSW and RF and SVR showing limited responses to the weighting strategy. In the overall comparison, PLSR remained the most stable model for cross-site SOC prediction in this study, and the improvement brought by SOC-SSW was mainly concentrated within the PLSR framework.

3.4. Statistical Robustness of the Principal Transfer

Bootstrap analysis supported the performance ranking shown in Table 5. The 95% confidence interval for direct-transfer bias remained entirely below zero, whereas the SOC-SSW bias interval included zero (Table 7), indicating that the systematic underestimation observed under direct transfer was substantially reduced. In paired comparisons, the confidence intervals for the SOC-SSW improvements in R^2, RMSE, MAE, and absolute bias relative to direct transfer did not cross zero. The reduction in paired absolute error also remained significant after Holm correction (adjusted $p = 0.0018$). Compared with the best observed 40% similarity-selection baseline, SOC-SSW likewise showed lower RMSE and MAE, with a significant paired absolute-error difference (adjusted $p = 0.0263$). These results indicate that the principal improvement was not attributable only to sampling variability in the target set.

Table 7. Prediction metrics and bootstrap 95% confidence intervals for the principal Gaofeng-to-Yachang PLSR transfer.

Method	R^2 (95% CI)	RMSE (95% CI)	MAE (95% CI)	Bias (95% CI)
Direct transfer	0.809 (0.739–0.853)	6.676 (5.670–7.594)	5.165 (4.418–5.934)	−3.842 (−4.814–−2.892)
Similarity selection 40%	0.811 (0.730–0.859)	6.627 (5.603–7.669)	4.910 (4.157–5.767)	−1.719 (−2.874–−0.594)
CORAL	0.542 (0.431–0.644)	10.322 (7.876–12.658)	6.789 (5.511–8.260)	−2.395 (−4.240–−0.644)
SOC-SSW	0.855 (0.789–0.898)	5.800 (4.893–6.652)	4.309 (3.627–5.031)	−0.945 (−1.967–0.068)

Note: RMSE, MAE, and bias are expressed in g kg^{-1}. Confidence intervals were calculated from 5000 paired bootstrap resamples.

3.5. Error Distribution Across SOC Levels and Source-Range Coverage

The error analysis showed that the effect of SOC-SSW was not uniform across the target SOC gradient (Table 8). Relative to direct transfer, the largest RMSE reductions occurred in the lowest and highest SOC quartiles, and the underestimation bias was also markedly reduced in these groups. In the second quartile, the main improvement was a reduction in bias, whereas the third quartile showed a moderate increase in RMSE and a shift toward overestimation. Of the 119 Yachang samples, 108 were within the Gaofeng SOC range, seven were below it, and four were above it. SOC-SSW still reduced RMSE and absolute bias for the within-range samples, showing that the overall improvement was not driven only by the small number of range-extrapolation cases. For the four samples above the source-domain maximum, underestimation was alleviated but remained pronounced, and this result should be interpreted cautiously because of the small subgroup size.

Table 8. PLSR prediction errors by measured Yachang SOC quartile.

SOC Group	SOC Range	n	Direct RMSE	Direct MAE	Direct Bias	SOC-SSW RMSE	SOC-SSW MAE	SOC-SSW Bias
Q1 (low)	4.26–11.21	30	4.456	3.796	−3.680	3.014	2.369	−0.407
Q2	11.35–17.30	30	3.899	3.183	−2.100	3.811	2.796	−0.109
Q3	17.66–29.34	29	6.025	4.945	−0.924	6.795	5.510	1.790
Q4 (high)	30.03–80.04	30	10.327	8.728	−8.568	8.075	6.602	−4.960

Note: SOC range and error metrics are expressed in g kg^{-1}.

3.6. Target-Spectrum Availability and Reverse Transfer

The target-spectrum availability analysis showed that using only 10 or 40 unlabeled Yachang spectra produced prediction accuracy broadly comparable to the complete-target condition (Table 9). The complete set yielded the lowest RMSE and smallest absolute bias, whereas the partial-target results varied slightly among random subsets, indicating that subset representativeness may be as important as sample number. In contrast, the exploratory Yachang-to-Gaofeng analysis produced negative R^2 values for all evaluated PLSR strategies (Table 10). Although CORAL had the lowest RMSE in this direction, none of the methods achieved acceptable reverse-transfer performance, and SOC-SSW did not improve the result. This contrast indicates that cross-site transfer was strongly direction-dependent.

Table 9. Sensitivity of Gaofeng-to-Yachang SOC-SSW to the number of unlabeled target spectra used for weighting.

Target Spectra	Fraction	R^2	RMSE (g kg^{-1})	MAE (g kg^{-1})	Bias (g kg^{-1})
10	8.4%	0.851 ± 0.010	5.886 ± 0.199	4.264 ± 0.137	−1.451 ± 0.447
40	33.6%	0.846 ± 0.003	5.991 ± 0.068	4.411 ± 0.066	−1.859 ± 0.147
119	100.0%	0.855	5.800	4.309	−0.945

Note: The 10-spectrum and 40-spectrum values are means ± SD from three random subsets. The 119-spectrum result is the complete-target principal experiment and was evaluated once.

Table 10. Exploratory Yachang-to-Gaofeng reverse-transfer performance using PLSR.

Strategy	R^2	RMSE	MAE	RPD	RPIQ	Bias
Direct transfer	−7.607	24.954	22.455	0.341	0.435	22.426
Similarity selection 40%	−2.527	15.973	11.707	0.533	0.680	−1.885
Similarity selection 50%	−1.108	12.349	9.936	0.690	0.880	−4.168
Similarity selection 60%	−2.669	16.293	12.514	0.523	0.667	−11.693
Similarity selection 70%	−31.951	48.826	45.946	0.174	0.223	45.409
CORAL	−0.503	10.429	8.001	0.817	1.042	2.395
SOC-SSW	−14.190	33.152	30.739	0.257	0.328	30.603

Note: Yachang (n = 119) was used as the labeled source domain and Gaofeng (n = 370) as the target domain. Gaofeng SOC labels were used only for final evaluation. RMSE, MAE, and bias are expressed in g kg^{-1}.

4. Discussion

4.1. Systematic Bias and the Role of Unlabeled Target-Domain Spectral Constraints

Direct source-domain transfer produced an underestimation bias, indicating that the spectral-SOC relationship established in Gaofeng Forest Farm could not be fully applied to Yachang Forest Farm. As shown in Figure 4, the raw reflectance of Yachang samples was generally higher than that of Gaofeng samples, but the Yachang SOC distribution was not concentrated in a low-SOC range and had a wider high-value range. This suggests that spectral differences between the two sites were not controlled only by SOC content, but may also have been influenced by soil color, mineral composition, particle-size composition, and scattering background [8,33]. However, these factors were not directly measured in this study. Although the mean spectral shapes appeared more similar after SG-SNV preprocessing, this visual comparison did not quantitatively demonstrate a reduction in source-target domain discrepancy. Local shifts nevertheless remained at several wavelengths. Therefore, unified preprocessing alone was insufficient to completely remove cross-site spectral distribution differences. For forest soil carbon monitoring, systematic underestimation may lead to underestimation of SOC levels in the target forest farm. SOC concentration predictions

cannot be directly interpreted as carbon-stock estimates without bulk density, soil depth, and other required variables. This systematic bias was the main reason for introducing unlabeled target-domain spectral constraints in this study. Although the direct-transfer R^2 in Yachang was numerically higher than the mean source-domain internal-validation R^2, the two values were calculated from datasets with different SOC variances and should not be interpreted as evidence of higher target-domain accuracy. RMSE, MAE, and bias provide more direct information for cross-dataset error comparison.

4.2. Roles of Target-Domain Spectral Similarity and SOC-Gradient Constraints

Similarity selection mitigated the bias of direct transfer, but its effectiveness depended on the selection ratio. When the top 40% most similar source-domain samples were retained, model bias was substantially reduced. Because the ratios were evaluated as an exploratory sensitivity analysis, the 40% result represents only the best observed setting for this source-target combination and should not be treated as an operationally optimized parameter. The response to the retention ratio was non-monotonic. Performance deteriorated at 50% and 60%, whereas the 70% setting recovered slightly but remained substantially worse than the 40% setting. This variability indicates that the number of retained samples alone does not determine transfer performance; the spectral and SOC-gradient representativeness of the retained source subset is also important. The advantage of SOC-SSW lies in not simply deleting samples but instead reducing the influence of dissimilar source-domain samples through similarity weights while maintaining the training contribution of different SOC levels through SOC stratification. This strategy used the distribution information provided by unlabeled target-domain spectra and avoided excessive concentration of source-domain samples within a local SOC range, thereby achieving higher prediction accuracy and lower systematic bias. The target-spectrum availability analysis further showed that useful weights could be constructed from 10 or 40 target spectra, although the complete target set produced the lowest RMSE and absolute bias. In practical terms, when SOC measurements are not yet available for a target forest farm, an existing source-forest spectral database can still be reused through the unlabeled spectra of the target domain, reducing the need for target-site SOC chemical analysis; however, target-soil sampling, sample preparation, and laboratory spectral acquisition are still required. This finding is consistent with studies on regional spectral libraries, local calibration, and memory-based learning that emphasize the importance of matching calibration samples with the target domain [34–38]. Compared with direct transfer and similarity selection, SOC-SSW required greater computation because 50 weighted-bootstrap PLSR submodels were trained rather than a single PLSR model. Nevertheless, the additional computation remained manageable for the present dataset on a standard laptop. Because wall-clock time depends strongly on hardware, the software environment, and parallel settings, no platform-independent runtime ratio is reported.

4.3. Performance Differences in CORAL and Regression Models, and Transfer Direction

CORAL did not improve target-domain prediction performance, suggesting that simply reducing the global covariance difference between source and target domains is not necessarily suitable for SOC spectral transfer. This outcome may be related to the relatively small dataset and to CORAL's emphasis on global second-order statistics rather than local SOC-related spectral structures. SOC prediction depends not only on consistency in global spectral distribution, but also on local spectral structures related to SOC concentration variation. If global alignment modifies these structures, it may reduce the model's ability to extract useful latent variables. The results for different models also showed that RF and SVR performed worse than PLSR in the target domain, possibly because they were more

likely to learn source-domain-specific nonlinear relationships. In contrast, PLSR is better suited to high-dimensional and multicollinear spectral data and can extract relatively stable spectral-SOC covariance information. It is therefore more suitable for combination with the SOC-SSW sample-weighting strategy. However, because RF and SVR did not improve under SOC-SSW, the present evidence supports mainly the SOC-SSW-PLSR combination rather than a model-independent benefit.

The failure of all evaluated methods in the reverse Yachang-to-Gaofeng transfer indicates that the difficulty was not specific to SOC-SSW. Reversing the domains reduced the labeled source set from 370 to 119 samples, which may have weakened the stability and representativeness of the calibration relationship. Direct reverse transfer already produced a strongly negative R^2, showing that the spectral–SOC relationship learned from Yachang was not transferable to Gaofeng before adaptation. Because the Gaofeng SOC range (6.665–64.535 g kg^{-1}) was fully contained within the Yachang range (4.260–80.040 g kg^{-1}), the failure cannot be attributed to SOC-range extrapolation alone and more likely reflects a direction-dependent mismatch in the spectral–SOC relationship. Similarity selection and CORAL reduced RMSE relative to direct transfer, but their negative R^2 values indicated that they could not recover a stable calibration relationship. For SOC-SSW, equalizing the contributions of SOC strata within the smaller and more heterogeneous source dataset may have increased the influence of sparsely represented high-SOC samples, contributing to the observed overprediction. This interpretation requires verification using larger multi-site datasets.

4.4. Limitations and Future Research

This study has several limitations. First, the current experiment used Gaofeng Forest Farm as the source domain and Yachang Forest Farm as the target domain; the applicability of the method to more target forest farms, different soil types, stand ages, and management backgrounds requires further validation. The exploratory reverse transfer was unsuccessful for all evaluated strategies, indicating strong direction dependence; the smaller Yachang source dataset (119 samples) may also have limited model stability. Second, this study used Vis-NIR spectra collected under laboratory conditions, where sample preparation and measurement environments were relatively stable. Therefore, the results cannot be directly equated with prediction performance under field in situ spectroscopy, UAV hyperspectral imagery, or satellite remote-sensing conditions. Third, the transfer strategy was constructed mainly from spectral distribution and the source-domain SOC gradient, without incorporating auxiliary variables such as soil texture, mineral composition, topography, and stand structure. Fourth, the numbers of PCA components, nearest neighbors, and SOC strata were fixed empirically. Although these settings were applied consistently without target-label-driven tuning, a comprehensive parameter-sensitivity analysis remains necessary. Fifth, a systematic comparison of alternative spectral preprocessing configurations and wavelength-exclusion schemes was not conducted. In addition, the stability of the Scenario C retention-ratio results under repeated perturbation or bootstrap resampling of the target spectral set was not evaluated; these issues should be examined in future work.

Spiking and many local-calibration methods were not included because they require measured target-domain SOC and therefore do not match the unlabeled-target setting considered here. Deep transfer-learning models were not evaluated because the datasets were relatively small and the study focused on transparent source-sample selection, covariance alignment, and instance weighting. These methods remain important benchmarks for future multi-site studies. Future research should extend the experiment to multiple source and target domains, examine how source-dataset size and representativeness affect transfer direction, and evaluate incremental or small-batch updating as target spectra be-

come available sequentially. Integrating sample weighting, feature-band selection, domain adaptation, and auxiliary soil variables may further improve stability and interpretability when target-domain physicochemical data are unavailable. Calibration-transfer methods and deep spectral models should also be compared under multi-source and multi-target conditions [39,40]. After independent multi-site validation, coupling the framework with digital soil mapping may support spatial assessment of SOC concentration.

5. Conclusions

This study proposed and evaluated the SOC-gradient-constrained spectral similarity weighting method (SOC-SSW) for the practical scenario in which a target forest farm lacks measured SOC samples and has only unlabeled Vis-NIR spectra. The main conclusions are as follows:

(1) The Gaofeng source-domain Vis-NIR spectral dataset showed stable SOC modeling feasibility. PLSR outperformed RF and SVR in source-domain internal validation and was therefore suitable as the main transfer-evaluation model in this study.

(2) Direct source-domain transfer captured the overall SOC variation trend in the Yachang target forest farm, but produced clear systematic underestimation. This indicates that, in cross-site applications, bias should be considered in addition to R^2 and RMSE when assessing SOC levels in the target forest farm.

(3) Unlabeled target-domain spectra provided effective distribution information for source-domain sample selection and weighting. The 40% similarity selection strategy mitigated direct-transfer bias, but simple sample selection depended on the SOC-gradient coverage of the retained source-domain samples. The retention ratios should therefore be interpreted as exploratory settings rather than target-label-free optimized parameters. Comparable accuracy was retained when weights were constructed from 10 or 40 target spectra, although the full target set produced the smallest absolute bias.

(4) SOC-SSW achieved the best cross-site prediction performance, with R^2, RMSE, RPD, and bias values of 0.855, 5.800 g kg^{-1}, 2.642, and -0.945 g kg^{-1}, respectively. Compared with direct source-domain transfer, SOC-SSW reduced RMSE by 13.12% and the absolute bias by 75.42%. Bootstrap and paired-error analyses confirmed the statistical robustness of this improvement in the Gaofeng-to-Yachang direction. However, the unsuccessful reverse transfer showed that the method is direction-dependent and requires a sufficiently large and representative source-domain dataset.

Overall, SOC-SSW combined with PLSR improved Gaofeng-to-Yachang SOC prediction and reduced systematic transfer bias without using target-domain SOC labels during model development. This benefit was conditional rather than universal, and broader multi-site validation is required before application to other plantation-forest settings.

Author Contributions: Conceptualization, Y.D. and Z.M.; methodology, Y.D. and Z.M.; software, Z.M.; validation, Z.M.; formal analysis, Z.M.; data curation, Z.M.; writing—original draft preparation, Z.M.; writing—review and editing, Y.D. and Z.M.; visualization, Z.M.; supervision, Y.D.; project administration, Y.D.; funding acquisition, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research study was funded by the Guangxi Key Research and Development Program (GuikeAB24010338 and GuikeAB25069340), the National Natural Science Foundation of China (32360374), and the Innovation Project of Guangxi Graduate Education (YCSW2024383 and YCSW2026782).

Data Availability Statement: The data used in this study are not publicly available as they are internal datasets of the research group. The data may be made available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation and revision of this manuscript, the authors used ChatGPT (GPT-5.6 Sol; OpenAI) for language editing, assistance with drafting and revising Python 3.9 scripts for additional analyses, and checking and organizing statistical results. The authors reviewed and verified all outputs and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

SOC	Soil organic carbon
Vis-NIR	Visible-near-infrared
SOC-SSW	SOC-gradient-constrained spectral similarity weighting
PLSR	Partial least squares regression
RF	Random forest
SVR	Support vector regression
SG	Savitzky–Golay
SNV	Standard normal variate
PCA	Principal component analysis
CORAL	Correlation alignment

References

1. Wiesmeier, M.; Urbanski, L.; Hobley, E.; Lang, B.; von Lützow, M.; Marin-Spiotta, E.; van Wesemael, B.; Rabot, E.; Ließ, M.; Garcia-Franco, N.; et al. Soil Organic Carbon Storage as a Key Function of Soils—A Review of Drivers and Indicators at Various Scales. *Geoderma* **2019**, *333*, 149–162. [CrossRef]
2. Minasny, B.; Malone, B.P.; McBratney, A.B.; Angers, D.A.; Arrouays, D.; Chambers, A.; Chaplot, V.; Chen, Z.-S.; Cheng, K.; Das, B.S.; et al. Soil Carbon 4 per Mille. *Geoderma* **2017**, *292*, 59–86. [CrossRef]
3. Li, S.; Viscarra Rossel, R.A.; Webster, R. The Cost-Effectiveness of Reflectance Spectroscopy for Estimating Soil Organic Carbon. *Eur. J. Soil Sci.* **2022**, *73*, e13202. [CrossRef]
4. Viscarra Rossel, R.A.; Brus, D.J. The Cost-Efficiency and Reliability of Two Methods for Soil Organic Carbon Accounting. *Land Degrad. Dev.* **2018**, *29*, 506–520. [CrossRef]
5. Long, M.; Yue, T.; Xu, Z.; Guo, J.; Luo, J.; Guo, X.; Zhao, X. Improved Soil Organic Carbon Prediction in a Forest Area by Near-Infrared Spectroscopy: Spiking of a Soil Spectral Library. *Forests* **2023**, *14*, 118. [CrossRef]
6. Viscarra Rossel, R.A.; Behrens, T.; Ben-Dor, E.; Chabrillat, S.; Demattê, J.A.M.; Ge, Y.; Gomez, C.; Guerrero, C.; Peng, Y.; Ramirez-Lopez, L.; et al. Diffuse Reflectance Spectroscopy for Estimating Soil Properties: A Technology for the 21st Century. *Eur. J. Soil Sci.* **2022**, *73*, e13271. [CrossRef]
7. Shin, S.K.; Lee, S.J.; Park, J.H. Prediction of Soil Properties Using Vis-NIR Spectroscopy Combined with Machine Learning: A Review. *Sensors* **2025**, *25*, 5045. [CrossRef] [PubMed]
8. Shepherd, K.D.; Ferguson, R.; Hoover, D.; van Egmond, F.; Sanderman, J.; Ge, Y. A Global Soil Spectral Calibration Library and Estimation Service. *Soil Secur.* **2022**, *7*, 100061. [CrossRef]
9. Bai, Z.; Xie, M.; Hu, B.; Luo, D.; Wan, C.; Peng, J.; Shi, Z. Estimation of Soil Organic Carbon Using Vis-NIR Spectral Data and Spectral Feature Bands Selection in Southern Xinjiang, China. *Sensors* **2022**, *22*, 6124. [CrossRef] [PubMed]
10. Zhang, T.; Li, Y.; Wang, M. Prediction of Soil Organic Carbon and Total Nitrogen Affected by Mine Using Vis-NIR Spectroscopy Coupled with Machine Learning Algorithms in Calcareous Soils. *Sci. Rep.* **2024**, *14*, 23634. [CrossRef] [PubMed]
11. Canero, F.M.; Rodríguez Galiano, V.; Aragón, D. Machine Learning and Feature Selection for Soil Spectroscopy: An Evaluation of Random Forest Wrappers to Predict Soil Organic Matter, Clay, and Carbonates. *Heliyon* **2024**, *10*, e30228. [CrossRef] [PubMed]
12. Ke, Z.; Ren, S.; Yin, L. Advancing Soil Property Prediction with Encoder-Decoder Structures Integrating Traditional Deep Learning Methods in Vis-NIR Spectroscopy. *Geoderma* **2024**, *449*, 117006. [CrossRef]
13. Seidel, M.; Hutengs, C.; Ludwig, B.; Thiele-Bruhn, S.; Vohland, M. Strategies for the Efficient Estimation of Soil Organic Carbon at the Field Scale with Vis-NIR Spectroscopy: Spectral Libraries and Spiking vs. Local Calibrations. *Geoderma* **2019**, *354*, 113856. [CrossRef]
14. Ng, W.; Minasny, B.; Jones, E.; McBratney, A. To Spike or to Localize? Strategies to Improve the Prediction of Local Soil Properties Using Regional Spectral Library. *Geoderma* **2022**, *406*, 115501. [CrossRef]

15. Padarian, J.; Minasny, B.; McBratney, A.B. Transfer Learning to Localise a Continental Soil Vis-NIR Calibration Model. *Geoderma* **2019**, *340*, 279–288. [CrossRef]
16. Shen, Z.; Ramirez-Lopez, L.; Behrens, T.; Cui, L.; Zhang, M.; Walden, L.; Wetterlind, J.; Shi, Z.; Sudduth, K.A.; Baumann, P.; et al. Deep Transfer Learning of Global Spectra for Local Soil Carbon Monitoring. *ISPRS J. Photogramm. Remote Sens.* **2022**, *188*, 190–200. [CrossRef]
17. Dai, L.; Xue, J.; Lu, R.; Wang, Z.; Chen, Z.; Yu, Q.; Shi, Z.; Chen, S. In-Situ Prediction of Soil Organic Carbon Contents in Wheat-Rice Rotation Fields via Visible Near-Infrared Spectroscopy. *Soil Environ. Health* **2024**, *2*, 100113. [CrossRef]
18. Viscarra Rossel, R.A.; Shen, Z.; Ramirez-Lopez, L.; Behrens, T.; Shi, Z.; Wetterlind, J.; Sudduth, K.A.; Stenberg, B.; Guerrero, C.; Gholizadeh, A.; et al. An Imperative for Soil Spectroscopic Modelling Is to Think Global but Fit Local with Transfer Learning. *Earth-Sci. Rev.* **2024**, *254*, 104797. [CrossRef]
19. Dorantes, M.J.; Fuentes, B.A.; Miller, D.M. Calibration Set Optimization and Library Transfer for Soil Carbon Estimation Using Soil Spectroscopy—A Review. *Soil Sci. Soc. Am. J.* **2022**, *86*, 879–903. [CrossRef]
20. Luce, M.S.; Ziadi, N.; Viscarra Rossel, R.A. GLOBAL-LOCAL: A New Approach for Local Predictions of Soil Organic Carbon Content Using Large Soil Spectral Libraries. *Geoderma* **2022**, *425*, 116048. [CrossRef]
21. Wang, Z.; Zhang, D.; Xu, X.; Lu, T.; Yang, G. Collaborative Utilization of Sentinel-1/2 and DEM Data for Mapping the Soil Organic Carbon in Forested Areas Based on the Random Forest. *Forests* **2024**, *15*, 218. [CrossRef]
22. Hau, N.-X.; Tuan, N.-T.; Trung, L.-Q.; Chi, T.-T. Estimation of Soil Organic Carbon Content Using Visible and Near-Infrared Spectroscopy in the Red River Delta, Vietnam. *Chemom. Intell. Lab. Syst.* **2024**, *255*, 105253. [CrossRef]
23. Yacomelo Hernández, M.J.; Ipanaqué Alama, W.; Montenegro, A.C.; Córdoba, O.J.; Castañeda Sanchez, D.; Vargas García, C.; Flórez Cordero, E.; Castillo Quezada, J.; Pacherres Herrera, C.; Prado-Castillo, L.F.; et al. Application of Vis-NIR Spectroscopy and Machine Learning for Assessing Soil Organic Carbon in the Sierra Nevada de Santa Marta, Colombia. *Sustainability* **2026**, *18*, 513. [CrossRef]
24. Lever, J.; Krzywinski, M.; Altman, N. Principal Component Analysis. *Nat. Methods* **2017**, *14*, 641–642. [CrossRef]
25. Kouw, W.M.; Loog, M. A Review of Domain Adaptation without Target Labels. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 766–785. [CrossRef] [PubMed]
26. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
27. Wang, Y.; Yin, K.; Hu, B.; Hong, Y.; Chen, S.; Liu, J.; Yang, L.; Peng, J.; Shi, Z. Ensemble and Transfer Learning of Soil Inorganic Carbon with Visible Near-Infrared Spectra. *Geoderma* **2025**, *456*, 117257. [CrossRef]
28. Wang, Y.; Yan, X.; Huo, R.; Zhao, L.; Peng, J.; Hong, Y.; Liu, J. Predicting Soil Organic Matter Using Corrected Field Spectra and Stacking Ensemble Learning. *Geoderma* **2025**, *460*, 117417. [CrossRef]
29. Das, B.; Chakraborty, D.; Singh, V.K.; Das, D.; Sahoo, R.N.; Aggarwal, P.; Murgaokar, D.; Mondal, B.P. Partial Least Square Regression Based Machine Learning Models for Soil Organic Carbon Prediction Using Visible-Near Infrared Spectroscopy. *Geoderma Reg.* **2023**, *33*, e00628. [CrossRef]
30. Tang, K.; Zhao, X.; Xu, Z.; Sun, H. A Stacking Ensemble Model for Predicting Soil Organic Carbon Content Based on Visible and Near-Infrared Spectroscopy. *Infrared Phys. Technol.* **2024**, *140*, 105404. [CrossRef]
31. Zuo, Z.; Tang, X.; Li, G.; Ma, Y.; Zhang, W.; Li, S. Inversion of Terrain Slope and Roughness with Satellite Laser Altimeter Full-Waveform Data Assisted by Shuttle Radar Topographic Mission. *Remote Sens.* **2021**, *13*, 424. [CrossRef]
32. Oberholzer, S.; Summerauer, L.; Steffens, M.; Ifejika Speranza, C. Best Performances of Visible-Near-Infrared Models in Soils with Little Carbonate—A Field Study in Switzerland. *SOIL* **2024**, *10*, 231–249. [CrossRef]
33. Wetterlind, J.; Viscarra Rossel, R.A.; Steffens, M. Diffuse Reflectance Spectroscopy Characterises the Functional Chemistry of Soil Organic Carbon in Agricultural Soils. *Eur. J. Soil Sci.* **2022**, *73*, e13263. [CrossRef]
34. Safanelli, J.L.; Hengl, T.; Parente, L.L.; Minarik, R.; Bloom, D.E.; Todd-Brown, K.; Gholizadeh, A.; Mendes, W.S.; Sanderman, J. Open Soil Spectral Library (OSSL): Building Reproducible Soil Calibration Models through Open Development and Community Engagement. *PLoS ONE* **2025**, *20*, e0296545. [CrossRef] [PubMed]
35. Miloš, B.; Bensa, A. Organic Carbon Estimation in a Regional Soil Vis-NIR Database Supported by Unsupervised Learning and Chemometrics Techniques. *Soil Adv.* **2024**, *2*, 100013. [CrossRef]
36. Wang, J.; Zou, Q.; Xu, B.; Feng, Z.; Yuan, H. Enhancing Soil Organic Matter Prediction via Deep Transfer Learning from Mid-Infrared Soil Spectral Library to In-Situ Field. *Comput. Electron. Agric.* **2026**, *241*, 111243. [CrossRef]
37. Tziolas, N.; Tsakiridis, N.; Ben-Dor, E.; Theocharis, J.; Zalidis, G. A Memory-Based Learning Approach Utilizing Combined Spectral Sources and Geographical Proximity for Improved Vis-NIR-SWIR Soil Properties Estimation. *Geoderma* **2019**, *340*, 11–24. [CrossRef]
38. Sun, Q.; Shi, P. Enhancing Proximal and Remote Sensing of Soil Organic Carbon: A Local Modelling Approach Guided by Spectral and Spatial Similarities. *Geoderma* **2025**, *457*, 117298. [CrossRef]

39. Chen, Y.; Shi, T.; Li, Q.; Yang, C.; Wang, Z.; Chen, Z.; Pan, X. Mapping Soil Properties in Tropical Rainforest Regions Using Integrated UAV-Based Hyperspectral Images and LiDAR Points. *Forests* **2024**, *15*, 2222. [CrossRef]
40. De Oliveira, K.M.; Gonçalves, J.V.F.; Furlanetto, R.H.; de Oliveira, C.A.; Mendonça, W.A.; Haubert, D.d.F.d.S.; Crusiol, L.G.T.; Falcioni, R.; de Oliveira, R.B.; Reis, A.S.; et al. Predicting Particle Size and Soil Organic Carbon of Soil Profiles Using VIS-NIR-SWIR Hyperspectral Imaging and Machine Learning Models. *Remote Sens.* **2024**, *16*, 2869. [CrossRef]