Computers and Electronics in Agriculture 242 (2026) 111354

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Domain-Invariant autoencoders and subset selection for Site-Specific clay prediction with small scale local data☆

Yongjing Wang , Abdul Mounem Mouazen *

Department of Environment, Faculty of Bioscience Engineering, Ghent University, 9000 Ghent, Belgium

ARTICLE INFO

Keywords:
NIR spectroscopy
Clay perdition
Domain adaptation
Autoencoders
Soil spectral libraries

ABSTRACT

The application of large soil spectral libraries (SSLs) for localized clay prediction is limited by domain shifts arising from soil moisture variability and instrumental differences. The study proposes a novel modelling and resampling framework for the prediction of clay content using a large SSL with dry samples and a local spectral library with wet samples. The core of the framework incorporated a domain-invariant autoencoder (DAE) with a maximum mean discrepancy (MMD) regularization to align spectral features between the dry-condition SSL and the wet-condition local dataset. A standard resampling method was then applied to the transformed SSLs to select an optimal subset of SSLs for site-specific calibration of clay content. Finally, partial least squares regression (PLSR) models were developed and validated using both a large local sample set and two subsets chosen via the standard resampling method (non-DAE-resampling) and proposed DAE-resampling methods. The performance of two spectrometers, namely, a AvaSpec-NIR (1000–2500 nm range) and a CompactSpec (305–1700 nm range) were evaluated. Results for the AvaSpec-NIR showed the DAE-resampling method achieved a prediction R^2 value of 0.74 with its corresponding PLSR model, comparable to the PLSR model trained on the large local calibration dataset ($R^2 = 0.78$), both significantly exceeding the non-DAE-resampling method ($R^2 = 0.53$). For the CompactSpec spectrometer, despite covered less clay-sensitive bands, the DAE-resampling method ($R^2 = 0.63$) still outperformed the non-DAE-resampling method ($R^2 = 0.48$) through enhanced cross-domain bias reduction. The DAE-resampling method also reduced the subset sampling computation time by over 50 % (AvaSpec-NIR: 30 s vs. 68 s; CompactSpec: 37 s vs. 86 s). While the DAE-resampling framework demonstrates potential for cost-effective soil clay content prediction with minimal local data, its performance remains sensitive to the NIR spectral range. Future work should explore alternative architectures and extend the framework to additional soil attributes. This study advances domain adaptation strategies in spectral analytics, offering a pathway to scalable soil monitoring in data-scarce situations.

1. Introduction

There is an increasing demand for affordable, high-quality and detailed spatial resolution data to support precision agriculture and address environmental issues like soil degradation. Monitoring soil clay content under these conditions is crucial, since clay affects nutrients holding capacity, water retention, and soil structure (Tümsavaş et al., 2019). However, obtaining accurate spatial clay content estimates on a large scale is challenging with traditional laboratory methods like hydrometer and pipette (Wang et al., 2015). Despite, traditional laboratory methods are accurate, they are labor-intensive, time-consuming, expensive and disposes chemical waste to the environment, which makes them unsuitable for a large amount of data analysis required in precision agriculture (Munnaf et al., 2019).

Diffuse reflectance spectroscopy in the visible and near infrared (vis–NIR) offers a rapid, non-destructive alternative for soil analysis under laboratory and in-situ conditions, enabling simultaneous calibration of multiple soil attributes, including clay content (Liu et al., 2024; Nocita et al., 2015; Omondiagbe et al., 2023; Stenberg et al., 2010). Many reports demonstrated the ability of vis-NIR spectroscopy for the prediction and mapping of clay content (Liu et al., 2024; Poggio et al., 2017; Tümsavaş et al., 2019). These results are not surprising thanks to the direct spectral responses the clay and clay minerology have in the NIR range (mostly above 2100 nm). However, there is a need for

☆ This article is part of a special issue entitled: 'Generative AI' published in Computers and Electronics in Agriculture.

* Corresponding author.
E-mail address: Abdul.Mouazen@UGent.be (A.M. Mouazen).

https://doi.org/10.1016/j.compag.2025.111354
Received 25 April 2025; Received in revised form 12 December 2025; Accepted 17 December 2025
Available online 24 December 2025
0168-1699/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

local site-specific calibrations in order to ensure accurate predictions. But, site-specific calibration is an expensive target to achieve, as it requires a large number of samples (+ 100) (Liu et al., 2024; Nawar et al., 2020; Tümsavaş et al., 2019), analyzed for clay content using the traditional laboratory methods (e.g., the hydrometer and pipette test). These requirements can be prohibitively expensive and time-consuming, particularly for those requiring prompt results or operating in regions with limited sampling resources.

Large soil spectral libraries (SSLs) provide a diverse range of soil types, mineralogical compositions, and corresponding vis-NIR spectra, which can potentially reduce the need for extensive local sampling. However, global calibration models using available SSLs do not always accurately predict soil attributes using local datasets, because spectral signals depend on local conditions such as mineral composition, organic matter, moisture content and climate (Canal Filho & Molin, 2022; Clairotte et al., 2016; Guerrero et al., 2014; Lobsey et al., 2017), few to mention among others.

This mismatch between the large-scale SSLs (the 'source domain') and the local, site-specific spectra (the 'target domain') is known as a domain shift (Pan & Yang, 2010). This shift is attributable to several factors, the most significant being the difference in measurement conditions. The SSLs typically consist of spectra from air-dried, ground, and sieved samples measured under laboratory conditions, whereas the local data is often collected in-situ from intact, fresh (wet) soils (Ackerson et al., 2017; Nawar et al., 2020). Soil moisture content (SMC) is a well-documented cause of domain shift that negatively impacts prediction accuracy in soil spectroscopy analyses. This occurs because moisture fundamentally alters the spectral signal, often increasing overall absorbance and masking key spectral features (Minasny et al., 2011; Mouazen et al., 2006a). Beyond moisture, this shift is compounded by instrumental differences, as SSLs and local data are often collected using different spectrometers with distinct optical probe designs and components (Ackerson et al., 2017; Nawar et al., 2020).

To address these challenges, the potential of using SSLs for local predictions has been investigated in several studies that select a subset of SSL to derive site-specific (local) calibrations. For example, LOCAL methods (Shenk et al., 1997), spectrum-based learning (Ramirez-Lopez et al., 2013) and resampling approach (Lobsey et al., 2017) have been explored for their ability to calibrate local models using minimal local data, selected from SSLs. However, because these methods were primarily developed for dry, laboratory-acquired spectra, techniques such as direct standardization (Chakraborty et al., 2019), piecewise direct standardization (Wang et al., 2002), and external parameter orthogonalization (Wijewardane et al., 2016) have been employed to mitigate the effects of moisture and environmental factors on soil spectra. However, these approaches require paired dry and wet spectra, a condition which is not met when calibrating site-specific models with minimal local data (in-situ wet [fresh] spectra) and SSL datasets (lab-based dry spectra). Moreover, the aforementioned subset selection methods are known to be time and computational demanding when applied to large SSLs. Furthermore, the practical application of in-situ soil sensing is often constrained by the cost, and the need for robust design of available spectrometers, factors that are affected by various spectrophotometer having different spectral ranges and optical structures. A robust calibration framework must therefore be stable and effective across these diverse instrumental conditions.

This highlights a critical research gap, which is addressed in the study by two primary objectives. The first is to propose and evaluate a novel domain-invariant autoencoder (DAE)-resampling framework to leverage SSLs for accurate, site-specific calibrations for clay prediction using in-situ fresh soil spectra with minimal labelled data. The second objective is to assess the robustness and practical performance of this framework for two distinct spectrometers with different wavelength coverages and optical designs. The core of framework integrates a DAE with a maximum mean discrepancy (MMD) regularization to align spectral features between the dry-condition SSL and the wet-condition local dataset. Subsequently, a standard resampling method (Lobsey et al., 2017) was applied to the transformed SSLs to select an optimal subset of SSLs for site-specific calibration of clay content. The study evaluated and compared the accuracy and efficiency of partial least squares regression (PLSR) models using both large site-specific calibration sample sets and two subsets of SSLs chosen via the proposed DAE-resampling and standard resampling without DAE adjustment methods (non-DAE-resampling).

2. Materials and methods

This study integrates a continental-scale SSL with a local fresh soil spectra set to develop a DAE-PLSR calibration model for the prediction of clay content. The SSL, derived from the LUCAS dataset (17,869 mineral soil samples), and a local dataset (197 labeled + 1000 unlabeled fresh soil samples) were preprocessed to harmonize spectral resolution and reduce noise. A DAE was trained on unpaired dry (SSL) and wet (local unlabeled) spectra to align their latent representations by minimizing domain shifts using a composite loss of mean squared error (MSE) and MMD. The resulting DAE-resampling framework combined this trained DAE with the resampling strategy of Lobsey et al. (2017) to select an optimal SSL subset in latent space for calibration using limited local target samples. Finally, PLSR models were trained on three different sets: (i) the DAE-resampling set, (ii) a traditional site-specific local (fresh) set, comprising a large volume of preprocessed local spectra, and (iii) a non-DAE-resampling set chosen by the conventional resampling method proposed by Lobsey et al., (2017) (non-DAE-resampling). The three models were validated using the same data set consisting of fresh spectra and measured clay content. An overview of the workflow is shown in Fig. 1, while the detailed materials and methods are presented in the following subsections.

2.1 . Data collection and processing

2.1.1. Soil spectral library (SSL)

The LUCAS dataset was employed as SSL in this study, originally comprising 19,967 surface soil samples (0–20 cm) collected in 2009 from various land use types across 25 European Union states (Tóth et al., 2013). The dataset includes 12 soil physicochemical properties (clay content inclusive), and the corresponding vis-NIR (400–2499.5 nm) spectral data at a 0.5 nm resolution. Spectral measurements were performed in the laboratory using a FOSS XDS Rapid Content Analyzer apparatus (FOSS Analytical, Hilleroed, Denmark), and clay content was determined with Robinson-Kohn pipette method (International Organization for Standardization, 2020). To focus on mineral soils, commonly associated with agricultural lands, and to maintain data integrity, samples lacking clay measurements or classified as organic soils were excluded, resulting in a refined SSL of 17,869 mineral soil samples for subsequent analysis.

2.1.2. Local soil sampling and spectral acquisition

Soil samples and in-situ fresh soil spectra were collected from 11 fields in three commercial farms, located in the west of the Flemish part of Belgium (Table 1). The in-situ spectral scanning and soil sampling campaigns were conducted between 21st of August 2023, and 11th of April 2024 (Table 1). These fields encompassed soil texture classes (Soil Science Division Staff, 2017) of sandy loam, silt loam and loam. The average gravimetric soil water content per field varied between 0.14 and 0.24 kg kg^{-1}. This study established two distinct local datasets to develop and validate the DAE framework:

- An unlabelled local in-situ fresh soil dataset (n = 1000) collected in-situ to capture the spectral characteristics of the target domain.
- A labelled local fresh soil dataset (n = 197) acquired under controlled laboratory conditions designed to simulate the in-situ sensor setups.

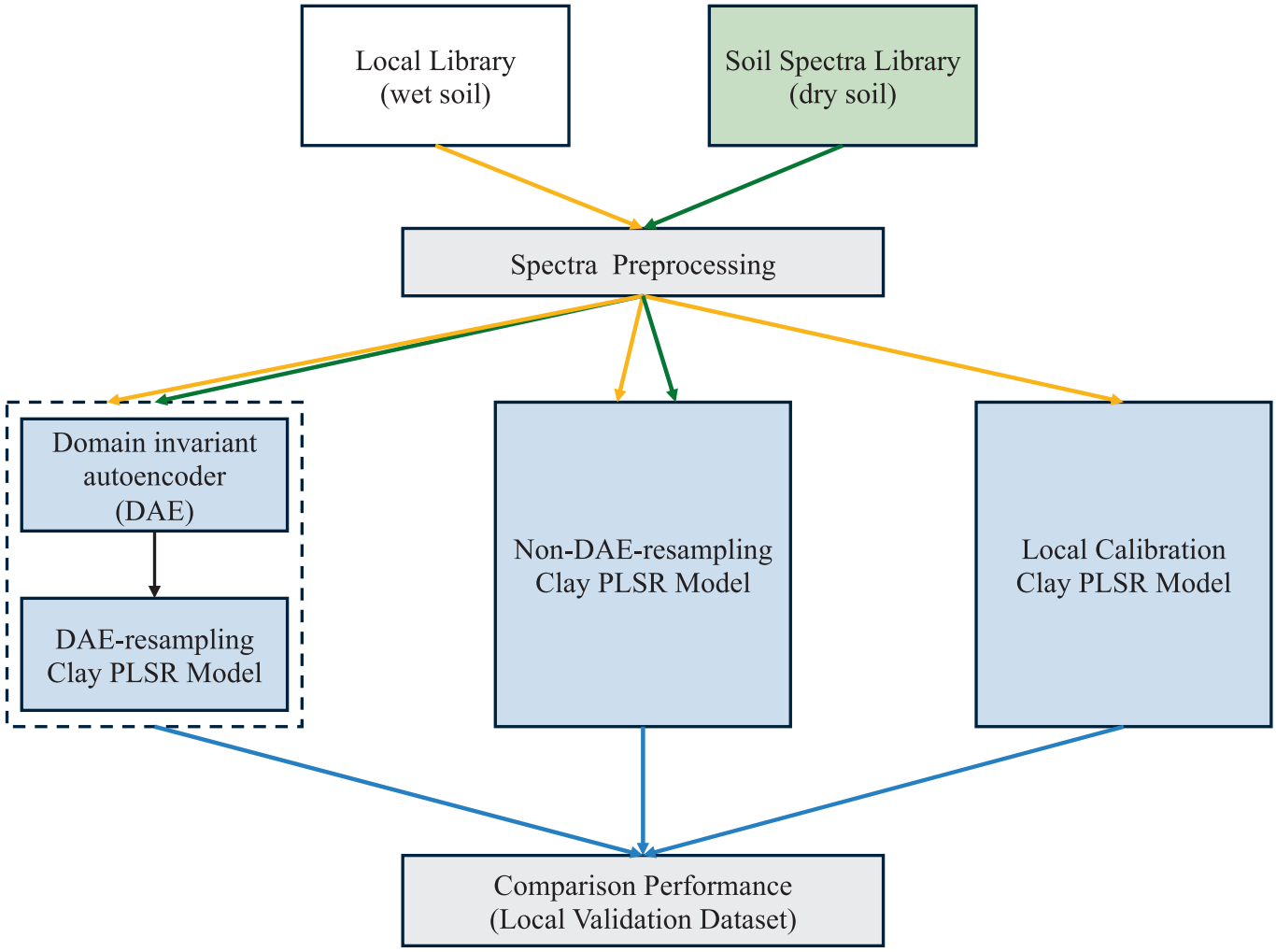


Fig. 1. Overview of the workflow followed in this study. PLSR denotes partial least squares regression.

Table 1
Field information and corresponding sampling number for this study.

Location	Date	Size (ha)	N	Soil Type	Texture	Clay (%)	Clay STD (%)	Silt (%)	Sand (%)	TOC (%)	SWC (kg kg^{-1})
50.99306, 2.53478	18/10/2023	12.5	15	Fluvisol	Loam	12	1.28	31	57	0.97	0.24
50.99093, 2.54005	22/08/2023	9.8	11	Fluvisol	Loam	23	6.50	45	33	Nan	0.31
51.01711, 2.54440	23/08/2023	20.4	30	Fluvisol	Loam	7	4.54	30	63	1.13	0.18
51.02848, 2.55840	21/08/2023	13.1	24	Fluvisol	Loam	7	3.64	11	82	2.16	0.14
50.91779, 3.75628	11/04/2024	7.5	29	Luvisol	Loam	7	1.89	9	84	1.14	0.22
50.91871, 3.75608	11/04/2024	2.9	15	Luvisol	Loam	7	0.72	15	78	2.03	0.21
50.98712, 3.75930	14/05/2024	2.8	15	Retisol	Silt Loam	7	3.64	10	83	2.16	0.20
51.11980, 3.50003	01/05/2024	8.5	24	Podzol	Sandy Loam	25	5.53	46	29	3.88	0.17
51.06582, 3.89062	10/04/2024	1.1	8	Podzol	Sandy Loam	33	0.46	52	15	6.76	0.18
51.07090, 3.89531	10/10/2023	3.0	12	Podzol	Sandy Loam	13	0.58	70	17	Nan	0.18
51.07462, 3.89467	10/04/2024	4.2	14	Podzol	Sandy Loam	13	2.71	70	17	1.21	0.23

N denotes the number of soil samples. SWC denotes the average soil water content (kg kg^{-1}) of the field. TOC denotes the average total organic carbon (%) of the field. STD is standard deviation. Clay, Silt, Sand denotes the average clay, silt and sand (%) of the field respectively.

2.1.2.1. In-situ spectral acquisition. To capture the true spectral characteristics of the target domain, the unlabelled dataset of 1000 spectra was collected in-situ. These spectra were acquired using two different in-situ platforms (Fig. 2):

- Platform 1 (Fig. 2A): An online multi-sensor soil platform (Mouazen, 2006) equipped with a CompactSpec spectrometer (Tec5 Technology for Spectroscopy, Germany), having a wavelength range of 305–1700 nm, a spectral resolution of 5 nm, and an output sampling resolution of 1 nm.
- Platform 2 (Fig. 2B): A portable multi-sensor penetrometer (Wang et al., 2025) equipped with an AvaSpec-NIR-2.5-HSC-EVO spectrometer (Avantes, The Netherlands), having a wavelength range of 1000–2500 nm with a 6.5 nm spectral and output sampling resolution.

This unlabelled dataset represents the full range of soil variability



Fig. 2. Two in-situ platform used in this study to acquire spectral data from fields: (A) the online multi-sensor soil platform (Mouazen, 2006), and (B) the portable multi-sensor penetrometer (Wang et al., 2025).

encountered in the field (e.g., variable moisture, texture, fertility, and physical conditions) and serves as the target domain for the DAE framework.

2.1.2.2. Labelled spectral acquisition. A total of 197 soil samples were obtained from the topsoil (15–25 cm depth). This depth was selected to correspond to the operational scanning depth of the in-situ platforms (Fig. 2) used to acquire the unlabelled target domain dataset. Based on the field sizes and sample numbers (Table 1), the sampling density for the labelled dataset was approximately 1–3 samples per hectare. It is important to note that this labelled dataset was not designed for field-specific calibration, but rather to create a spectrally diverse local dataset for evaluating the different modelling frameworks. During sample preprocessing, stones and organic residents were manually removed. Each sample was divided into two portions: one was immediately scanned with two vis–NIR spectrometers in the laboratory, while the other part was subjected to conventional analysis of clay content using the Robinson-Kohn pipette method following International Organization for Standardization (2020). Scanning the labelled samples under controlled, simulated conditions was a deliberate methodological

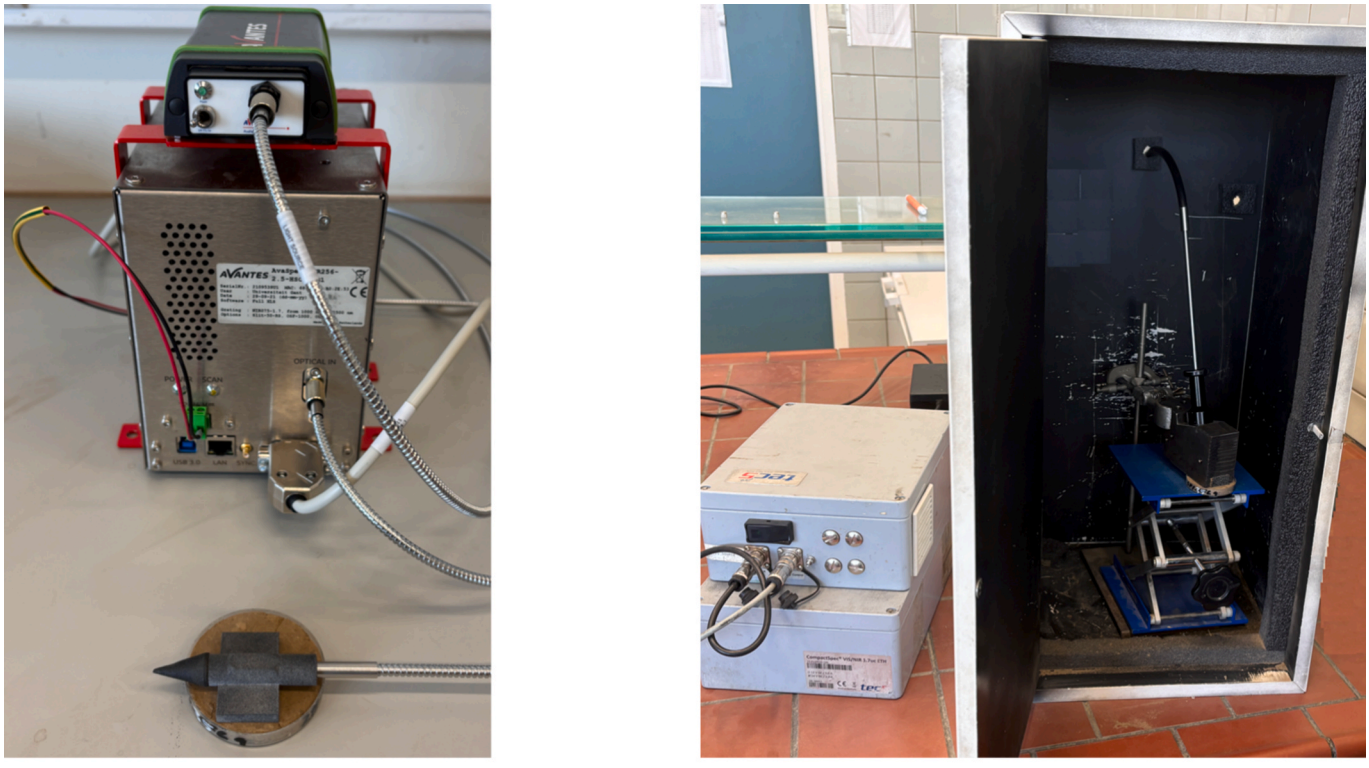


Fig. 3. Two experimental setups used in this study to acquire spectral data from soil samples: (A) the CompactSpec spectrometer, and (B) the AvaSpec-NIR spectrometer.

choice. This approach was chosen to: (i) guarantee a precise 1:1 match between each spectrum and its corresponding laboratory-measured clay content, eliminating potential field labelling errors, and (ii) focus specifically on the primary domain shifts due to instrument differences and SMC by controlling other random confounding field variables like vibration or ambient light.

Two vis-NIR spectrometers systems used in the laboratory (Fig. 3) were designed to replicate the optical configurations of the in-situ platforms. The first system (Fig. 3.A) comprised a Y-shaped reflection probe coupled with a 20 W halogen light source and a CompactSpec spectrometer. The Y-shaped reflection probe tip was integrated into a metal module housing containing a light condensing lens, while scanning was carried out within a black box to avoid effects of the ambient light. The second system (Fig. 3.B) used a similar Y-shape reflection probe connected to a 10 W halogen light source and an AvaSpec-NIR-2.5-HSC-EVO spectrometer. In this setup, the probe tip was inserted into a 3D printed cylinder with a safari window (6 mm in diameter), covered by a 2 mm thick sapphire lens and a 45° mirror positioned behind the window. Two adjacent wings were designed to maintain a fixed 2 mm distance from the sample surface and minimize ambient light interference during spectral scanning. For spectral acquisition, fresh soil samples were uniformly placed into Petri dishes and levelled with a spatula to enhance the signal-to-noise ration. Each sample underwent five spectral scans by each spectrometer. The averaged spectrum was retained for analysis to minimize random noise and potential inconsistencies from the manual scanning process, thereby providing a more stable and representative spectral signature for each sample. This approach is similar to procedures used by SSL (LUCAS) (Nocita et al., 2014). During the measurement process, a PTFE disc with approximately 99 % reflectance was used to calibrate the sensor every 30 min.

2.1.3. Spectra preprocessing

Spectra acquired from the LUCAS SSL were initially resampled via linear interpolation to conform to the wavelength resolution and range of the AvaSpec-NIR spectrometer, and then subjected to the same preprocessing steps. The raw spectra from AvaSpec-NIR spectrometer (1000–2500 nm, 6.5 nm resolution) were preprocessed by trimming noisy segments at both spectral edges, yielding a refined spectral range of 1055 to 2323 nm. These raw spectra were subsequently processed using Savitzky-Golay smoothing and a standard normal variate (SNV) transformation (Barnes et al., 1989) to mitigate noise and improve the signal-to-noise ratio, resulting in smoothed spectra. Similarly, for the spectra from the CompactSpec spectrometer (305–1700 nm, 1 nm output sampling resolution), an initial resampling to the instrument's native 5 nm spectral resolution was performed using linear interpolation, followed by splice correction at 1045 nm (Mouazen & Ramon, 2009) and the removal of noisy edge segments, yielding a spectral range of 500 to 1595 nm as raw spectra. These raw spectra were subsequently smoothed using Savitzky-Golay filtering and SNV transformation. The LUCAS SSL spectra intended for comparison with CompactSpec data were resampled to the corresponding wavelength resolution and range prior to undergoing the same spectra preprocessing procedures to ensure consistency.

2.2. Domain invariant autoencoder (DAE)

Previous resampling approach proposed by Lobsey et al. (2017) was applied to dry samples in both the SSL and the local data set. To extend this approach to local wet-condition spectra without relying on laboratory clay measurements, a DAE is introduced to align spectral features between the dry-condition SSL and the wet-condition local dataset in an unsupervised manner. The DAE employs a symmetric encoder-decoder neural network architecture to learn compressed, unsupervised representations of data. The encoder maps high-dimensional inputs into a latent space, while the decoder reconstructs the original input from this latent representation. In this work, the challenge involves processing unpaired dry spectra (from the SSL dataset) and wet spectra (from the local dataset), where a critical objective is to ensure the latent features are not biased toward domain-specific characteristics of either dataset. To this end, the proposed DAE incorporated a MMD loss (Gretton et al., 2012), applied to the encoder's latent representations. By minimizing the MMD loss between latent distributions of the two domains, the model enforces alignment, ensuring the latent space captures domain-invariant features essential for cross-domain generalization.

The DAE was implemented in Python 3.10 using PyTorch 2.4.1 with graphic procession unit (GPU) acceleration. The encoder comprises three down sampling blocks, each utilizing a 1D convolutional layer (kernel size = 9, stride = 2) with a SiLU activation function. The final convolutional output (z map) is compressed via a max pooling layer to produce a 1D latent vector. The decoder mirrors this structure, with three transpose convolutional layers using identical configurations to up sample the z map, followed by a fully connected layer to reconstruct the original spectra (Fig. 4). During training, the model minimizes a composite loss function that simultaneously ensures accurate spectral reconstruction and enforces domain invariance in the latent space. This composite loss comprises two components. The first component is the MSE loss, which penalizes wavelength-level discrepancies between the original and reconstructed spectra. The second component is the MMD loss, computed using a radial basis function (RBF) kernel, which aligns the latent distributions of the SSL (dry) and local (wet) spectra. The final training loss is defined as equation (1):

$$\text{loss} = 0.7 \times \text{MSE loss} + 0.3 \times \text{MMD loss} \quad (1)$$

This weighting formulation was chosen to prioritize the accurate reconstruction of spectral features (MSE loss, 0.7 weight), which is essential for the model's predictive utility, while simultaneously enforcing domain alignment (MMD loss, 0.3 weight) to mitigate the dataset shift.

The model was trained on unpaired 1D spectral data, consisting of 17,869 dry spectra from the LUCAS SSL dataset and 1000 wet spectra from the local unlabelled dataset. A random split allocated 70 % of the SSL and local unlabelled data for training and the remaining 30 % for validation. The raw spectra, prior to smoothing and SNV preprocessing, were used for model development, as the DAE with convolutional layers effectively mitigates the influence of noise and baseline shifts, eliminating the need for additional spectra preprocessing. Training was performed for a maximum of 1000 epochs, with early stopping implemented after 30 consecutive epochs without improvement. In addition, the best model was saved during training to avoid overfitting. The composite loss function was minimized using an AdamW (Loshchilov & Hutter, 2019) optimizer with a dynamic learning rate schedule (ReduceLROnPlateau) and an initial learning rate of 0.01. Two distinct DAE models were trained on two cases:

1) A model trained using the LUCAS SSL dry spectra, combined with unlabeled fresh local spectra from the CompactSpec spectrometer, and
2) A model trained using the LUCAS SSL dry spectra, combined with unlabeled fresh local spectra from the AvaSpec-NIR spectrometer.

The reconstruction performance of the DAE was evaluated using MMD, root mean square error (RMSE) and Pearson correlation coefficient (PCC) metrics.

2.2.1. DAE-resampling algorithm

To address the challenge of domain variance between dry spectra in the SSL and wet spectra in the local dataset, the DAE-resampling algorithm was developed in this study by integrating the previously described DAE with the resampling algorithm (Lobsey et al., 2017). Once the optimal DAE model was obtained through training, both the SSL and local target spectra were projected into a domain-invariant latent space via the encoder component of DAE. The DAE-resampling

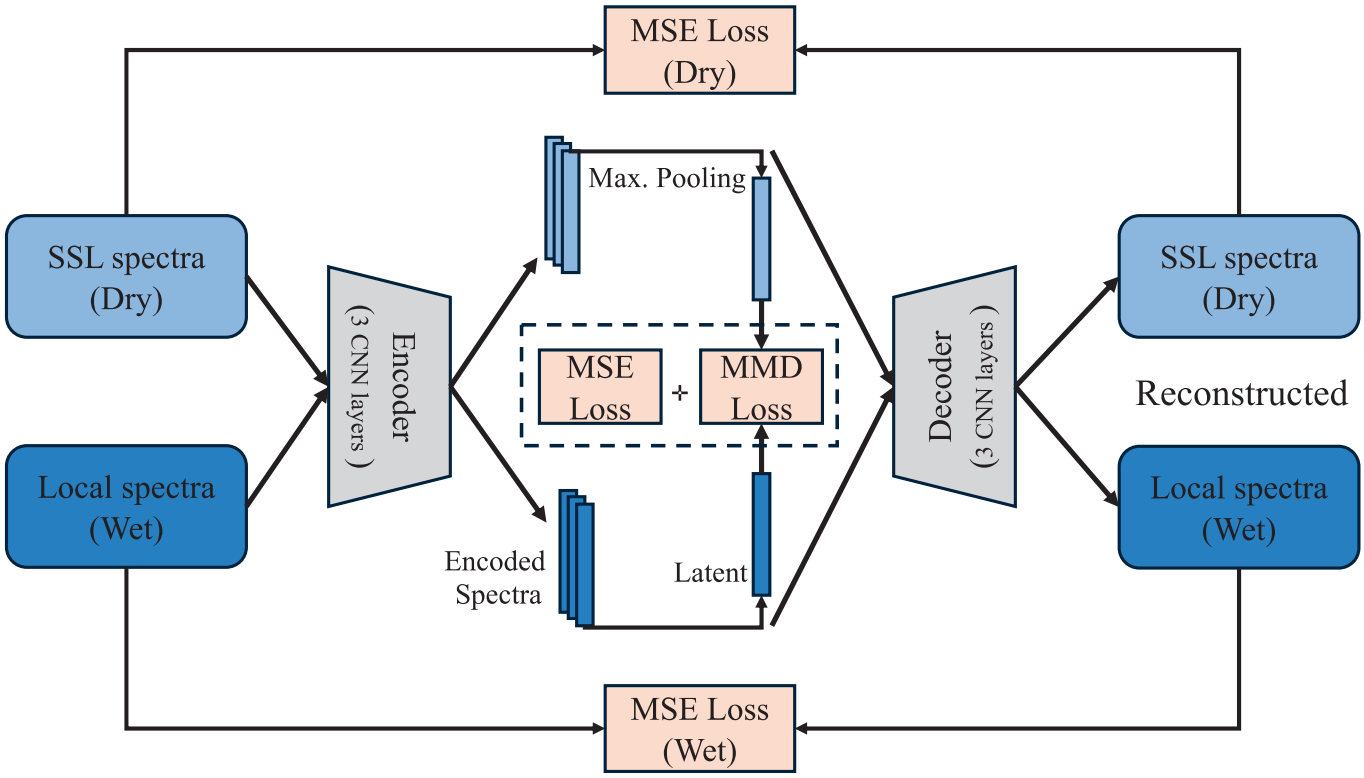


Fig. 4. Proposed architecture of the domain-invariant autoencoder (DAE) featuring the maximum mean discrepancy (MMD) based regularization. MSE denotes the mean squared error and SSL denotes the soil spectral library (LUCAS spectral library was used in this study).

algorithm subsequently identified the most relevant SSL samples for the local target dataset by following the standard resampling methodology but operating within the latent space rather than the full spectral dimension. This hybrid approach mitigates domain shifts between two datasets, while reducing computational costs through dimensionality reduction, without compromising predictive accuracy. Based on the parameter configurations for the full spectral dimension for standard resampling algorithm (e.g., non-DAE-resampling) proposed by Lobsey et al. (2017), the DAE-resampling algorithm, adopted as a novel approach in this study, implements the following steps:

1. Project all SSL and local target spectra into the latent space using the trained DAE.
2. Create S as a subset of the SSL, with S initially containing all SSL data.
3. Select a training set of k pairs from S by random sampling without replacement.
4. Develop a PLSR model with 8 components using the selected training set to predict clay content.
5. Validate the PLSR model against the local target dataset (L) by calculating the root mean square error (RMSE) and assign this RMSE to the selected k training samples.
6. Repeat steps 2–4 for n iterations, where $n = 17869 \times (b/k)$, with $b = 40$ (the target average number of selections per sample, a parameter adopted from the original resampling method by Lobsey et al. (2017)). Although the training set (k pairs) is sampled without replacement in step 2, the samples are returned to S before each repeat, and the RMSE from step 4 is cumulatively assigned to each selected sample over the n repeats.
7. Rank SSL samples in S based on cumulative RMSE, standardized by the recorded frequency of selection. Remove the top 10 % of samples with the highest standardized RMSE.
8. Repeat steps 2 to 7 until the size of the dataset S is reduced to k.

2.3. Experimental setup

2.3.1. Data splitting

The full local labelled dataset comprised 197 paired samples of clay content and wet spectra obtained from the CompactSpec and AvaSpec-NIR spectrometers. To ensure consistent evaluation of the DAE-resampling algorithm against the non-DAE-resampling algorithm (Lobsey et al., 2017) and traditional site-specific calibration models, the full local dataset was first partitioned using the Kennard-Stone (KS) method (Kennard & Stone, 1969) into a local calibration dataset (137 samples) and a local validation dataset (60 samples). Subsequently, a local target dataset of 27 samples was selected from the local calibration dataset using the KS method. The KS algorithm was applied to ensure that the calibration, target and validation sets were spectrally representative of the combined 11 fields. This ensures a robust and unbiased comparison between the three modelling approaches. This local target dataset simulated a scenario, in which only a limited amount of local data was available and was used to select a subset of SSL samples via both the DAE-resampling and non-DAE-resampling algorithms, while the local calibration dataset was employed to train the traditional site-specific clay prediction model. The validation dataset remained the same and was not included in all calibration procedures (including subset selection and model training) to ensure unbiased performance comparisons. Spectra from the CompactSpec and AvaSpec-NIR spectrometers were treated as two independent scenarios; in each case, the spectra were paired with the clay content in parallel and used for model calibration.

2.3.2. Clay prediction and evaluation

To assess the effectiveness of the proposed DAE-resampling method in selecting SSL subsets for site-specific calibration with limited local target data, PLSR models were trained and evaluated under three configurations:

1 Local calibration dataset (137 fresh samples) with preprocessed spectra: This set represents the standard site-specific calibration approach using a large number of samples from local datasets, as widely documented in vis-NIR spectroscopy studies.
2 Local target dataset (27 samples) + SSL subset (110 samples) with raw spectra: The samples of the subset selected using the non-DAE-resampling algorithm were combined with the local target dataset making a new dataset with 137 samples for model training.
3 Local target dataset (27 samples) + SSL subsets (110 samples) with raw spectra: The samples of the subset selected via the DAE-resampling algorithm were combined with the local target dataset making a new dataset with 137 samples for model training. All spectra were transformed into the latent space with the DAE encoder prior to analysis.

The SSL subset size of 110 samples was chosen to ensure that the total size of the resampling datasets (110 SSL + 27 local = 137 samples) precisely matches that of the local calibration dataset (n = 137). This equalization of sample size is crucial for a direct and unbiased comparison of model performance. Furthermore, the impact of varying the SSL subset size (i.e., the resampling ratio) was not re-evaluated in this study. This decision is supported by the original findings of Lobsey et al. (2017), who demonstrated that prediction accuracy was not significantly influenced by the size of the selected SSL subset across a wide range of 50 to 600 samples. By controlling for total dataset size and leveraging this established finding, our configuration isolates the evaluation to the primary innovation (DAE framework) of this study.

The predictive performance of the models was evaluated using the coefficient of determination (R^2), RMSE, the residual prediction deviation (RPD), and the ratio of performance to interquartile range (RPIQ). The configurations of the experimental setups used to evaluate and compare the three PLSR models are also visually depicted in Fig. 5.

3. Results and discussion

3.1. Soil and spectral data of the two datasets

Table 2 summarizes the clay content statistics for both the SSL (LUCAS) and local (local full labelled) datasets. The SSL dataset exhibited an average clay content of 0.19 kg kg^{-1} with a standard deviation (STD) of 0.13 kg kg^{-1}, which is consistent with the large diversity of soil types, texture, and land use conditions across multiple European countries, from which the SSL (LUCAS) dataset was collected (Tóth et al., 2013). In contrast, the local fresh dataset, collected from 11 fields in three farms in the Flemish region of Belgium, had smaller average clay content of 0.14 kg kg^{-1} and STD of 0.08 kg kg^{-1}, indicating lower variability. It is worth mentioning in particular the large deviation of the maximum value between the SSL and the local dataset, indicating

Table 2
Summary statistics for the soil clay content (kg kg^{-1}) of the global soil spectra library (SSL) and local datasets.

	Count	Mean	STD	Min.	25 %	50 %	75 %	Max.
SSL	17,869	0.19	0.13	0.01	0.08	0.17	0.26	0.79
Local	197	0.14	0.08	0.03	0.07	0.12	0.17	0.43

STD denotes standard deviation; SSL denotes soil spectral library (LUCAS); Local denotes the full local labelled dataset.

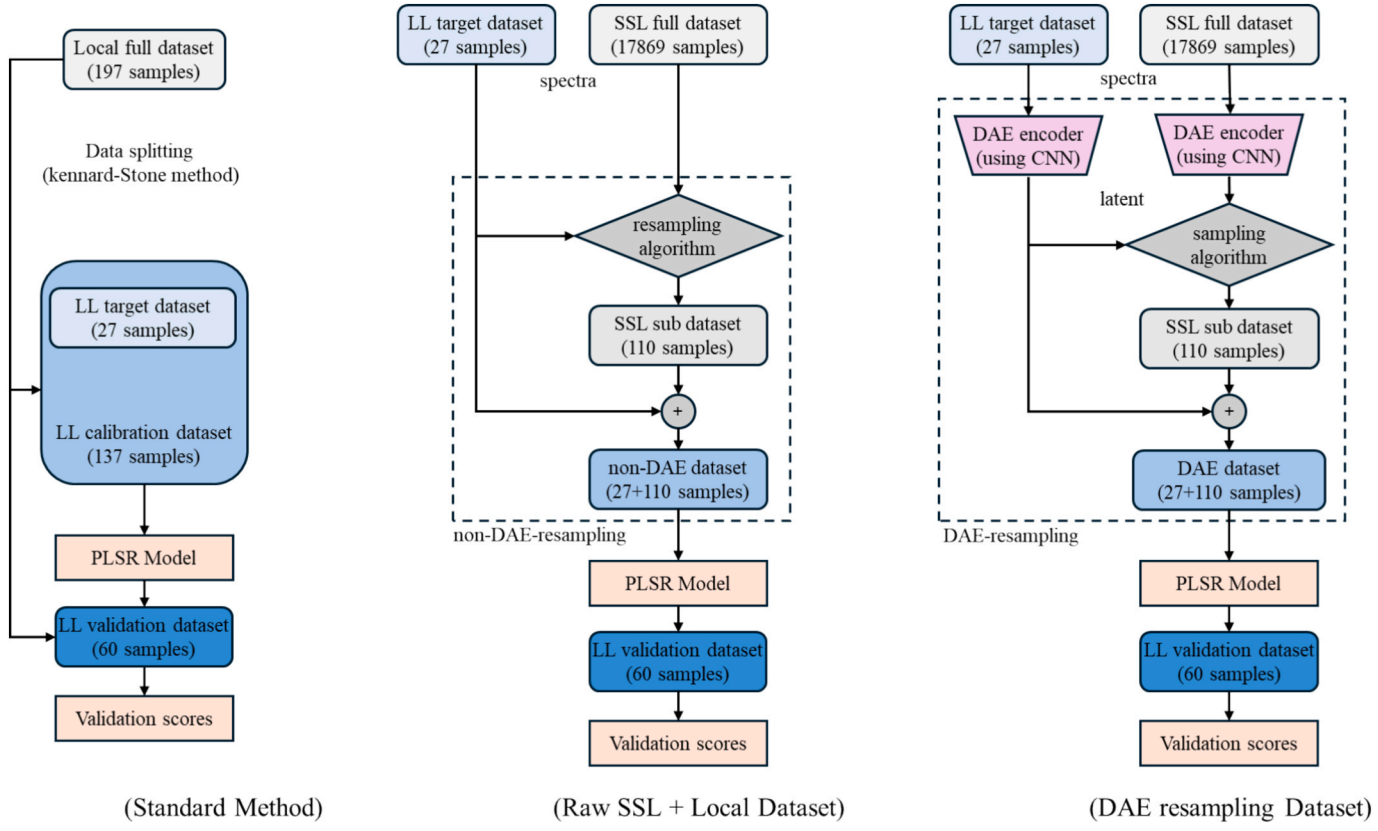


Fig. 5. Development and training workflows of the three PLSR models compared in this study. (A) Training workflow for the local calibration dataset, including the data-splitting schema for the full local labelled dataset, which is consistently applied across all workflows. (B) Training workflow that incorporates an SSL subset selected by the non-DAE resampling algorithm along with the local target dataset. (C) Training workflow incorporating an SSL subset selected by the proposed DAE-resampling algorithm alongside the local target dataset. SSL denotes the LUCAS soil spectral library; LL denotes local dataset; DAE denotes domain-invariant autoencoder; CNN denotes convolutional neural network layer; PLSR denotes partial least squares regression.

a challenge for the unsupervised learning approach adopted in this work (e.g., the non-DAE-resampling algorithm to select a subset of SSL for local model calibration).

Fig. 6 presents the mean raw reflectance spectra, with STDs, derived from both the SSL and local datasets employed in this study. As described in the Materials and Methods section, the range of raw spectra of the SSL dataset were trimmed and resampled to correspond with that of each spectrometer device (AvaSpec-NIR and CompactSpec) used locally. Three significant absorbance features, observed at approximately 1400, 1900, and 2200 nm, were identified. The spectral differences between the SSL and local dataset are more pronounced around 1400 and 1900 nm, which are strongly associated with free water OH features in the second and the first overtones of OH (Mouazen et al., 2006a). The absorption feature near 2200 nm is likely attributable to the Al-OH absorption present in clay content (Viscarra Rossel et al., 2006). Since the soil was dry in the SSL dataset, the spectral reflectance was higher than that of the fresh soil samples in the local dataset.

3.2. The performance of DAE reconstruction

Two distinct DAE models were trained separately using raw spectra from SSL and local unlabeled datasets, corresponding to the AvaSpec-NIR and CompactSpec spectrometers, as previously described. Fig. 7 presents both the mean raw input spectra and the mean reconstructed spectra for the two models. In both cases, the deviation between the raw and reconstructed spectra was minimal, suggesting effective feature extraction and accurate spectral reconstruction by the DAE models. This conclusion was further supported by the low MSE values between the input and reconstructed spectra along with a mean PCC of 0.99 for both datasets reported in Table 3 (a PCC value of 1 indicates perfect spectral reconstruction).

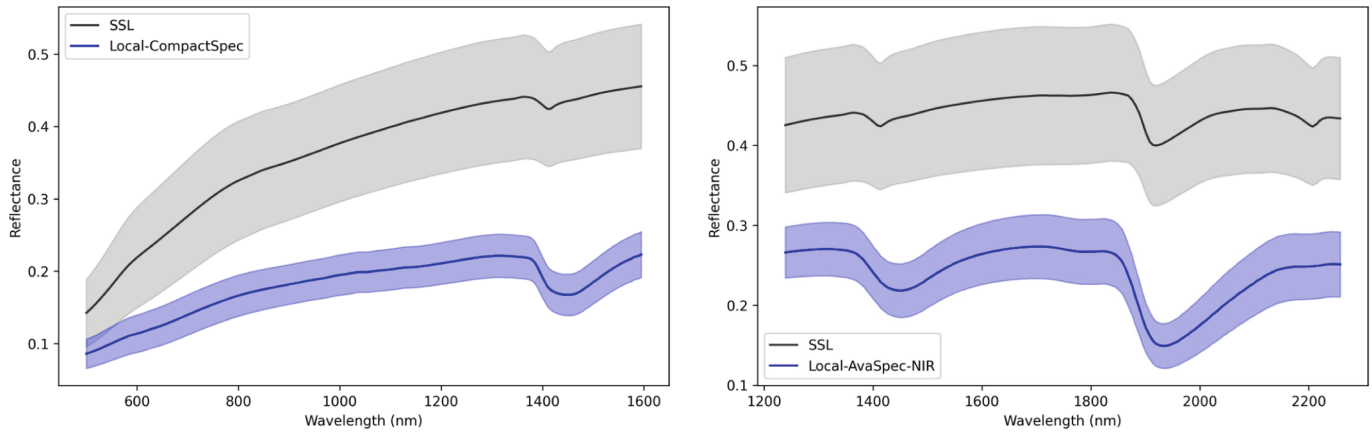
Domain disparity analysis via MMD indicated significant distributional differences between the SSL and local datasets, with values of 1.114 for the AvaSpec-NIR's DAE and 1.113 for the CompactSpec's DAE (Table 3). However, the MMD values computed for the encoded latent representations decreased dramatically to 3.93×10^{-6} for AvaSpec-NIR and 7.15×10^{-7} for CompactSpec. This marked reduction demonstrated that the DAE encoders effectively project the input spectra into a shared latent space since lower MMD values signify increased similarity between the distributions of the two datasets (Gretton et al., 2012), thereby substantially reducing domain variance. Collectively, these findings confirmed that both DAE models successfully mitigated inter-domain discrepancies between the SSL and local datasets while preserving essential spectral characteristics, validating their efficacy in cross-domain spectral feature alignment.

3.3. Evaluation of DAE-resampling dataset for clay perdition

Three PLSR models for clay prediction were trained using distinct calibration datasets: one with the large local (site specific) calibration dataset (137 pairs of clay content and soil spectra), one with the subset of SSL selected by the non–DAE-resampling plus local target dataset (27 pairs), and one with the subset of SSL selected by DAE-resampling plus local target dataset (27 pairs). Although the training metrics of these models are not directly comparable due to the different calibration datasets, their performance was evaluated using the identical local validation dataset of 60 samples.

For the AvaSpec-NIR spectrometer (Table 4), the DAE-resampling PLSR model demonstrated a comparable predictive accuracy to the traditional local model (Fig. 8.A), achieving a validation R^2 of 0.74, RMSE of 0.039 kg kg^{-1} and RPIQ of 3.32 (Table 4), which are close to the local model's validation R^2 of 0.78, RMSE of 0.035 kg kg^{-1} and RPIQ of 3.68. In contrast, the non-DAE-resampling model exhibited a lower validation R^2 of 0.53 and PRIQ of 2.48, a poor performance attributed to the domain variance between the local dataset (wet spectra) and SSL (dry spectra), which is the issue addressed by the DAE-resampling method by integrating MMD regularization within the DAE to minimize distributional differences. For the CompactSpec spectrometer (Table 4), the local model again showed superior performance (validation $R^2 = 0.74$, RPIQ = 3.60), while both DAE-resampling and non-DAE-resampling models exhibited reduced predictive capacity (validation $R^2 = 0.63$ and 0.48, RPIQ = 3.05 and 2.55, respectively). Notably, the DAE-resampling model outperformed non-DAE-resampling model (Fig. 8.B), underscoring its relative robustness despite domain disparities. Furthermore, the DAE-resampling method exhibited significantly lower computational demands for subset sampling from the SSL dataset compared to the non-DAE-resampling approach (AvaSpec-NIR: 30 s vs 68 s; CompactSpec: 37 s vs 86 s), thus enabling enhanced calibration efficiency.

The AvaSpec-NIR spectrometer (1000–2500 nm) covers all three significant bands (1400, 1950 and 2200 nm) observable in Fig. 6, with the ~2200 nm while being less associated and affected by soil moisture content, it is the band that has the strongest association with clay and clay minerology (Viscarra Rossel & Behrens, 2010). In contrast, the CompactSpec spectrometer encompasses only the ~1400 nm region, which overlaps with water absorption band at the second overtone of OH (Mouazen et al., 2006a). Consequently, the DAE-resampling method encountered greater challenges in mitigating domain variance between



(A) (B)

Fig. 6. The trimming of the range of the raw spectra obtained from the LUCAS soil spectra library (SSL) and the local Belgium datasets. (A) displays the spectra acquired using the CompactSpec spectrometer, whereas (B) shows the spectra recorded with AvaSpec-NIR spectrometer.

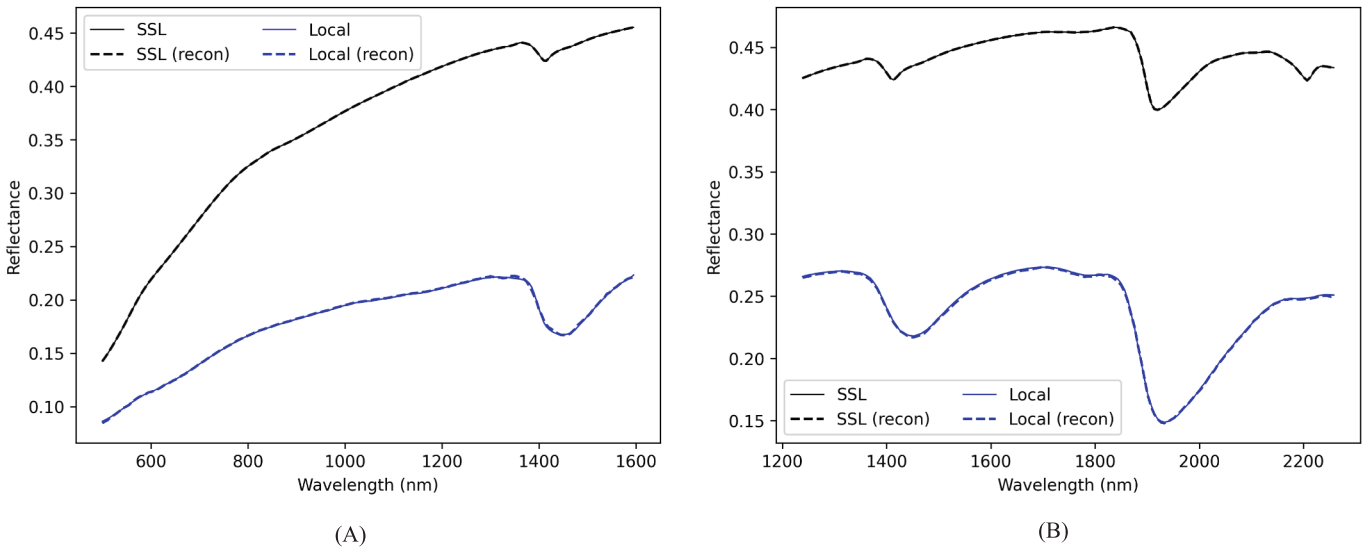


Fig. 7. Trimmed mean raw spectra (solid line) from the LUCAS soil spectra library (SSL, black) and local Belgium datasets (blue), alongside the corresponding mean reconstructed spectra (dashed line) by means of domain-invariant autoencoder (DAE) shown for the CompactSpec (A) and AvaSpec-NIR (B) spectrometers.

Table 3
Maximum mean discrepancy (MMD) values quantifying the domain differences between the LUCAS soil spectra library (SSL) and local unlabelled datasets for both the raw spectra and the latent representations encoded by the domain-invariant autoencoder (DAE) models. The Table also shows the mean squared error (MSE) and Pearson correlation coefficient (PCC) values between the raw input spectra and their corresponding reconstructions produced by DAE models for both the LUCAS SSL and local Belgium unlabelled dataset.

	MMD (latent)	MMD (raw)	SSL		Local	
			MSE	PCC	MSE	PCC
AvaSpec-NIR	1.114	3.93×10^{-6}	3.19×10^{-6}	0.99	1.18×10^{-6}	0.99
CompactSpec	1.113	7.15×10^{-7}	1.17×10^{-6}	0.99	7.70×10^{-7}	0.99

local (wet soil) and SSL (dry soil) spectra for the CompactSpec spectrometer, resulting in poorer model performance compared to that obtained for the AvaSpec-NIR spectrometer. This disparity highlights the critical role of spectral range and moisture-insensitive bands in addressing dataset discrepancies for robust clay prediction.

3.4. Discussions

Mouazen et al. (2006b) studied the effect of spectral range on the prediction accuracy of key soil physical and chemical attributes under fresh and dry conditions using the local calibration scheme of data collected in Belgium and Northern France. Authors reported the dry condition models outperformed the corresponding fresh soil models, and that the larger wavelength range of 350–2500 nm spectrometer outperformed the shorter wavelength range spectrometer (350–1750 nm) for all studied parameters. However, for the fresh soil models the performance was not clear, while the full range spectral model overperformed the short range spectrometer mainly for properties having spectral signatures beyond 2000 nm, e.g., soil organic carbon and organic nitrogen. This finding can also apply for the clay content in this studfym as clay has significant spectral signatures around 2200 and 2350 nm, which are covered by the AvaSpec-NIR spectrometer explaining the high accuracy of clay prediction achieved in this work compared to that of the CompactSpec spectrometer for fresh samples, regardless if DAE is applied or not.

Kuang et al. (2011) further demonstrated that the accuracy of vis-NIR predictions can vary widely between farms depending on the concentration range, showing large differences in independent validation R^2 of 0.89, 0.74 and 0.14 among different farms in Europe for organic

Table 4
Calibration and validation statistics for soil clay content predictions in kg kg^{-1} using three partial least squares regression (PLSR) models for the AvaSpec-NIR spectrometer and the CopmatSpec spectrometer, developed using three datasets of: 1) the local calibration dataset (Local), 2) a subset selected from the LUCAS soil spectral library (SSL) by the domain-invariant autoencoder (DAE)-resampling algorithm, and 3) a subset sampled by the non-DAE-resampling algorithm.

Methods	N	Time (s)	Calibration				Validation			
			R^2	RMSE	RPD	RPIQ	R^2	RMSE	RPD	RPIQ
AvaSpec-NIR										
Local	6	nan	0.81	0.041	2.29	3.56	0.78	0.035	2.16	3.68
DAE-resampling	5	30	0.75	0.068	2.00	2.53	0.74	0.039	1.94	3.32
non-DAE-resampling	5	68	0.70	0.057	1.83	2.94	0.53	0.043	1.45	2.48
CompactSpec										
Local	8	nan	0.83	0.038	2.42	3.92	0.74	0.039	1.96	3.60
DAE-resampling	6	37	0.81	0.046	2.31	1.94	0.63	0.046	1.66	3.05
non-DAE-resampling	7	86	0.51	0.071	1.43	2.11	0.48	0.055	1.39	2.55

N denotes the optimal number of latent variables (components) in the final PLSR model. 'nan' indicates that the computation time is not applicable for the local method, as it does not involve a resampling algorithm. R^2 denotes the coefficient of determination, RMSE, the root mean square error, RPD, the residual prediction deviation, and RPIQ the ratio of performance to interquartile range.

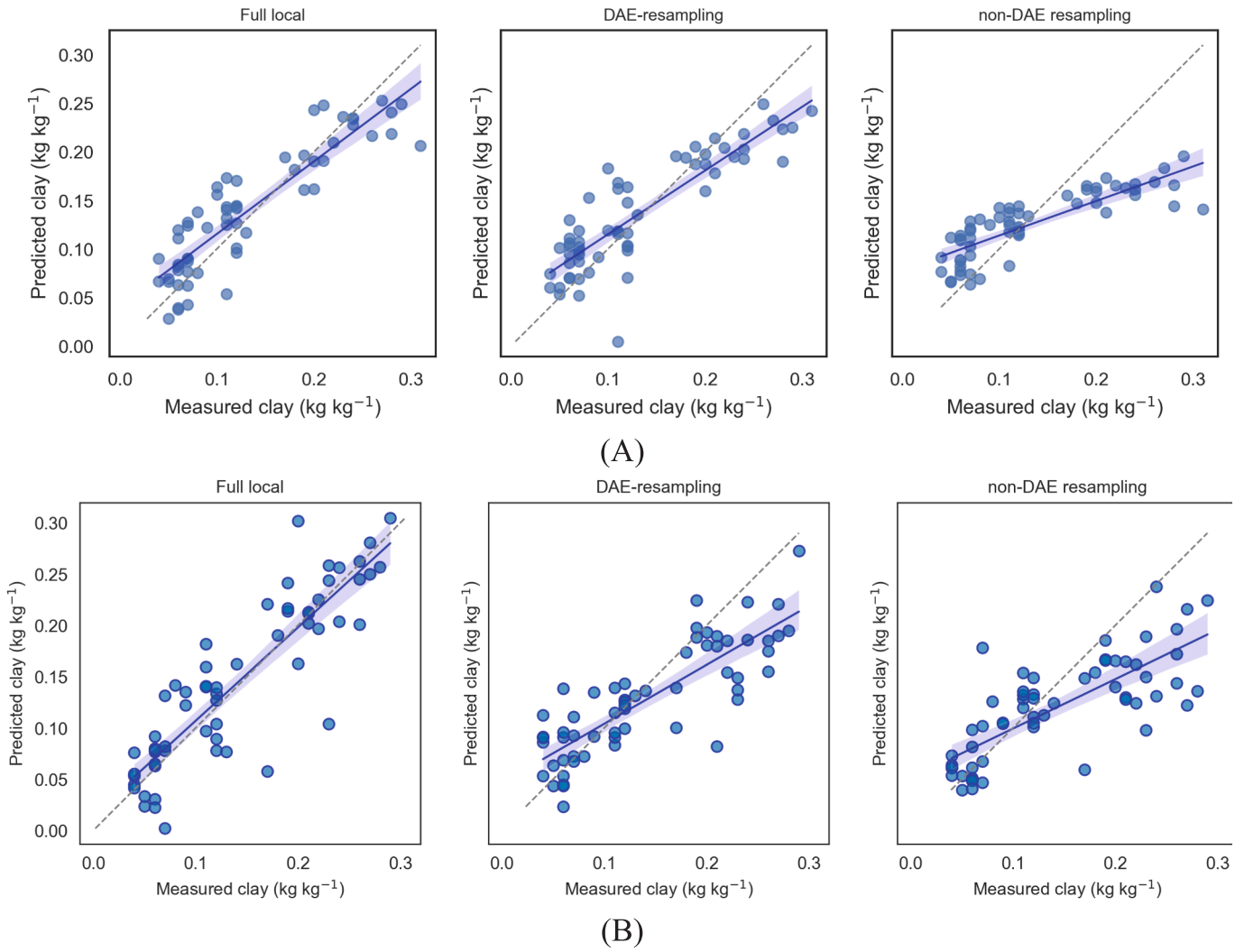


Fig. 8. Scatter plots of predicted versus measured soil clay content using the full local dataset, domain-invariant autoencoder (DAE)-resampling dataset and non-DAE resampling dataset for the AvaSpec-NIR spectrometer (A) and the CopmatSpec spectrometer (B). The blue circles represent the individual validation samples. The solid blue line indicates the linear regression line at the 95% confidence interval (shaded area), and the dashed grey line represents the 1:1 reference line.

carbon. They suggested that narrow concentration ranges undermine the robustness of the calibration, while overly heterogeneous datasets inflate the measurement error beyond acceptable thresholds. Although these findings pertain to organic carbon, the same principle holds for other soil properties, including clay content. This point is highly relevant to our study. As shown in Table 2, our local dataset possesses a narrower clay concentration range (0.03–0.43 kg kg^{-1}) than the SSL (0.01–0.79 kg kg^{-1}). Relying solely on a small local target set (n = 27) would likely result in a model with poor robustness. The DAE-resampling framework addresses this by allowing the final calibration to leverage the much wider concentration range of the SSL, thereby improving model robustness. Collectively, previous investigations underscore three principal pitfalls in vis-NIR spectroscopic soil analysis: (1) calibration models based on excessively narrow concentration ranges lack robustness; (2) models exploiting broader spectral windows generally outperform those constrained to limited wavelength ranges; and (3) drying soil samples typically enhances model performance by mitigating moisture-induced spectral variability. Kuang et al. (2015) demonstrated that an online PLSR calibration of clay achieved a validation R^2 value of 0.75, using 241 locally acquired fresh spectra with a similar device setup to that of the CompactSpec spectrometer (350–1700 nm) used in the present study. Consistent with these findings, the proposed DAE-resampling model exhibited superior performance using the AvaSpec-NIR spectrometer (1000–2500 nm), achieving a validation R^2 of 0.74, whereas the CompactSpec (350–1700 nm) enabled a lower R^2 of 0.63, due to the spectral range not capturing the clay spectral features. These results underscore that calibration models covering the high end spectral range (2000 – 2500 nm), and spanning broader spectral ranges generally outperform those restricted to narrower wavelength windows, ignoring the high NIR spectral range end.

On top of that moisture content was reported to have significant effects on prediction of soil attributes (Stenberg et al., 2010). Researchers utilised successfully different methods to remove this negative effect, including drying of samples, the spectra transferormation and correction for orthogonality (Ji et al., 2016). The spectra transofrmation method included direct standardisation (DS) and piecewise direct standardisation (PDS), whereas the correction for orthogonality included external parameter orthogonalization (EPO) and orthogonal signal correction (OSC). Mirzaei et al. (2022) demonstrated that simple air-drying of soil samples elevates clay-prediction accuracy demonstrated by increasing in R^2 from 0.45 on wet spectra to 0.89 on dry spectra when using full-range vis-NIR data (350–2500 nm). Liu et al. (2024) similarly observed that a dry-soil model (R^2 = 0.56) outperformed its wet-soil counterpart (R^2 = 0.36) using micro-NIR spectra (2000–2450 nm). To address moisture-induced variability without requiring sample drying, Mirzaei et al. (2022) developed an EPO

preprocessing routine applied to 330 spectra spanning moisture levels of 8, 16, 24, 30 and 36 %, which led to improved wet-sample model R^2 from 0.45 (pre-EPO) to 0.79 (post-EPO). However, all studies using fresh soil spectra or adopting algorithms to remove the effect of soil moisture (e.g., EPO) needed to use the full local dataset (+100 samples), which is an expensive procedure and does not fulfil the requirement when only a limited number of samples is available. In the present study, the DAE-resampling framework enabled the utilization of a limited number of fresh samples (20 samples) enabling successful calibration of clay content, thanks to the adoption of reconcile domain differences between spectra from LUCAS SSL and local dataset while preserving key clay-sensitive features. On the AvaSpec-NIR spectrometer (1000–2500 nm), DAE-resampling model achieved a validation R^2 of 0.74, compared to 0.53 for the non-DAE variant. A similar trend was observed with the CompactSpec instrument (350–1700 nm), where DAE-resampling model increased R^2 from 0.48 (non-DAE-resampling model) to 0.63.

While standard domain adaptation techniques such as EPO or DS are effective for removing moisture effect, they rely on paired wet and dry spectra to calculate the correction matrix. Their implementation typically requires large datasets (e.g., >100 samples) spanning multiple moisture levels, along with corresponding laboratory reference analyses for parameter optimization (Minasny et al., 2011). In practice, acquiring such paired datasets for every new target site is labour-intensive, costly, and time consuming. In contrast, the DAE-resampling framework proposed here operates in an unsupervised manner, requiring no paired samples. By aligning the distributions of the source (SSL) and target (local) domains in a latent space, it achieves comparable prediction accuracy to local calibration without the strict data requirements of EPO or DS, making the DAE-resampling approach more practical for rapid, cost-effective clay prediction in the field. Furthermore, compared to simple subset-selection methods (e.g., non-DAE resampling), the DAE-resampling ensures that the selected spectra are not just statistically similar, but semantically aligned with the target domain's moisture conditions, leading to significantly more robust predictions.

Employing the DAE-resampling framework obviates the need to air-dry soil samples and reduces both the cost and logistical burden of clay modelling by relying on a small set of fresh soil spectra. The practical advantage of this framework is the significant reduction in required local number of samples. The traditional local model required 137 labelled samples to achieve high accuracy ($R^2 = 0.78$ for AvaSpec-NIR). The DAE-resampling framework achieved a comparable result ($R^2 = 0.74$) using only 27 labelled target samples, an 80 % reduction in laboratory analysis costs. Thus, given that the traditional local model utilized a large number of samples (137 samples in this work), it is advised to use the DAE-resampling framework for AvaSpec-NIR spectrometer to predict clay, as efficiency can be improved significantly, cost of clay analysis can cut down considerably, while preserving a similar level of accuracy. It is critical to reiterate that while the DAE-resampling method consistently outperformed the non-DAE approach, it did not surpass the traditional, data-intensive local calibration model in the case of the CompactSpec spectrometer ($R^2 = 0.63$ vs. 0.74). This limitation underscores the critical role of the spectral range, as the DAE's ability to find domain-invariant features for clay prediction is severely hampered when clay's sensitive bands (e.g., ~2200 nm) are not present. However, in case it is necessary to use the CompactSpec spectrometer, one should balance between accuracy and cost in choosing between the local model (more expensive, and requires higher soil sampling and analysis time) and the DAE-resampling proposed for the first time in this work.

A key consideration is the translation of this study's experimental design to a practical field application. Our methodology used a KS-split on an existing labelled dataset to simulate the 'local target' and 'local validation' sets for a robust model comparison. In a real-world scenario, however, a practitioner would first scan a field densely with an in-situ platform to acquire a large, spatially comprehensive unlabelled spectral dataset. A selection algorithm, such as KS, could then be applied to these unlabelled spectra to identify a small, spectrally representative subset of locations for physical sampling and lab analysis. This 'scan-first, sample-later' approach would ensure both spectral and spatial coverage, providing the minimal 'local target' data needed for the DAE-resampling framework. While our current labelled dataset was too sparse for a field-by-field validation, future work should focus on validating this practical workflow, likely integrating geostatistical principles to optimize the selection of target samples.

4. Conclusions

This study demonstrates the potential of integrating DAE with a MMD regularization and a standard resampling algorithm to select a SSL subset, enabling the extraction of a domain-invariant latent representation that mitigates spectral disparities between SSL (dry-condition) and local (wet-condition) datasets. The PLSR for the AvaSpec-NIR spectrometer, which spans moisture diagnostic bands (~1450 and 1950 nm) and clay sensitive band (2200 nm), after the DAE-resampling subset selection achieved a prediction accuracy of clay content comparable to the model trained on the large local calibration dataset. Both models significantly outperformed the corresponding non-DAE-resampling-PLSR model. Although, this comparable prediction performance after implementing DAE-resampling diminished for the CompactSpec spectrometer, having shorter NIR range up to 1700 nm, the DAE-resampling PLSR model ($R^2 = 0.63$) still surpassed the non-DAE-resampling model ($R^2 = 0.48$) by more effectively reducing cross-domain biases. These results demonstrated the viability of site-specific clay prediction with minimal local calibration data using the DAE-resampling framework combined with PLSR, emphasizing the critical role of spectrometer spectral range in model performance, particularly under moisture-induced domain shifts. The findings further highlight the potential of domain-invariant deep learning techniques to enable cost-effective NIR-based soil property prediction in scenarios where large-scale local sampling is infeasible.

Despite these promising results, some limitations remain. First, the DAE-resampling method exhibited sensitivity to spectrometer spectral range, as evidenced by its deteriorated prediction quality with CompactSpec data. Its performance remains sensitive to the NIR spectral range. Specifically, the framework's utility is limited when crucial clay-sensitive spectral bands (>2000 nm) are absent from the spectrometer's range. Second, while the MMD-regularized autoencoder successfully aligned cross-domain spectral features, exploration of alternative neural architectures or supplementary regularization strategies could enhance robustness. Finally, extending this framework to other soil properties (e.g., organic carbon, nutrient concentrations) may necessitate adjustments for both the latent space design and subset selection criteria, underscoring the need for domain-specific adaptations.

CRediT authorship contribution statement

Yongjing Wang: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Abdul Mounem Mouazen:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors acknowledge the financial support received from WHEAT-WATCHER project, funded by the European Union under the Horizon Europe program and Swiss State Secretariat for Education, Research and Innovation (SERI), grant agreement No. 101156480.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2025.111354.

Data availability

Data will be made available on request.

References

Ackerson, J.P., Morgan, C.L.S., Ge, Y., 2017. Penetrometer-mounted VisNIR spectroscopy: Application of EPO-PLS to in situ VisNIR spectra. Geoderma 286, 131–138. https://doi.org/10.1016/j.geoderma.2016.10.018.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. Appl. Spectrosc. 43 (5), 772–777. https://doi.org/10.1366/0003702894202201.

Canal Filho, R., Molin, J.P., 2022. Spatial distribution as a key factor for evaluation of soil attributes prediction at field level using online near-infrared spectroscopy. Front. Soil Sci. 2. https://doi.org/10.3389/fsoil.2022.984963.

Chakraborty, S., Li, B., Weindorf, D.C., Morgan, C.L.S., 2019. External parameter orthogonalisation of eastern European VisNIR-DRS soil spectra. Geoderma 337, 65–75. https://doi.org/10.1016/j.geoderma.2018.09.015.

Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N.P.A., Bernoux, M., Barthès, B.G., 2016. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. Geoderma 276, 41–52. https://doi.org/10.1016/j.geoderma.2016.04.021.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012. A kernel two-sample test. J. Mach. Learn. Res. 13 (null), 723–773.

Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R.A., Maestre, F.T., Mouazen, A.M., Zornoza, R., Ruiz-Sinoga, J.D., Kuang, B., 2014. Assessment of soil organic carbon at local scale with spiked NIR calibrations: Effects of selection and extra-weighting on the spiking subset. Eur. J. Soil Sci. 65 (2), 248–263. https://doi.org/10.1111/ejss.12129.

International Organization for Standardization. (2020). Soil quality—Determination of particle size distribution in mineral soil material—Method by sieving and sedimentation (No. ISO 11277:2020). https://www.iso.org/standard/69496.html.

Ji, W., Li, S., Chen, S., Shi, Z., Viscarra Rossel, R.A., Mouazen, A.M., 2016. Prediction of soil attributes using the chinese soil spectral library and standardized spectra recorded at field conditions. Soil Tillage Res. 155, 492–500. https://doi.org/10.1016/j.still.2015.06.004.

Kennard, R.W., Stone, L.A., 1969. Computer aided Design of Experiments. Technometrics 11 (1), 137–148. https://doi.org/10.1080/00401706.1969.10490666.

Kuang, B., Mouazen, A.M., 2011. Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms. Eur. J. Soil Sci. 62 (4), 629–636. https://doi.org/10.1111/j.1365-2389.2011.01358.x.

Kuang, B., Tekin, Y., Mouazen, A.M., 2015. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. Soil Tillage Res. 146, 243–252. https://doi.org/10.1016/j.still.2014.11.002.

Liu, J., Munnaf, M.A., Mouazen, A.M., 2024. Micro-Near-Infrared (Micro-NIR) sensor for predicting organic carbon and clay contents in agricultural soil. Soil Tillage Res. 242, 106155. https://doi.org/10.1016/j.still.2024.106155.

Lobsey, C.R., Viscarra Rossel, R.A., Roudier, P., Hedley, C.B., 2017. Rs-local data-mines information from spectral libraries to improve local calibrations. Eur. J. Soil Sci. 68 (6), 840–852. https://doi.org/10.1111/ejss.12490.

Minasny, B., McBratney, A.B., Bellon-Maurel, V., Roger, J.-M., Gobrecht, A., Ferrand, L., Joalland, S., 2011. Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. Geoderma 167–168, 118–124. https://doi.org/10.1016/j.geoderma.2011.09.008.

Mirzaei, S., Darvishi Boloorani, A., Bahrami, H.A., Alavipanah, S.K., Mousivand, A., Mouazen, A.M., 2022. Minimising the effect of moisture on soil property prediction accuracy using external parameter orthogonalization. Soil Tillage Res. 215, 105225. https://doi.org/10.1016/j.still.2021.105225.

Mouazen, A.M., 2006. Soil Survey Device (World Intellectual Property Organization. International Bureau Patent No. WO/2006/015463.

Mouazen, A.M., Ramon, H., 2009. Expanding implementation of an on-line measurement system of topsoil compaction in loamy sand, loam, silt loam and silt soils. Soil Tillage Res. 103 (1), 98–104. https://doi.org/10.1016/j.still.2008.09.012.

Mouazen, A.M., Karoui, R., De Baerdemaeker, J., Ramon, H., 2006. Characterization of soil water content using measured visible and near infrared spectra. Soil Sci. Soc. Am. J. 70, 1295–1302.

Munnaf, M., Nawar, S., Mouazen, A.M., 2019. Estimation of Secondary Soil Properties by Fusion of Laboratory and On-Line measured Vis–NIR Spectra. Article 23 Remote Sens. (Basel) 11 (23). https://doi.org/10.3390/rs11232819.

Loshchilov, I., Hutter, F., 2019. Decoupled Weight Decay Regularization No. arXiv: 1711.05101). arXiv. https://doi.org/10.48550/arXiv.1711.05101.

Nawar, S., Abdul Munnaf, M., Mouazen, A.M., 2020. Machine Learning based On-Line Prediction of Soil Organic Carbon after Removal of Soil Moisture effect. Article 8 Remote Sens. (Basel) 12 (8). https://doi.org/10.3390/rs12081308.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biol. Biochem. 68, 337–347. https://doi.org/10.1016/j.soilbio.2013.10.022.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Wetterlind, J., 2015. Chapter four - Soil Spectroscopy: an Alternative to Wet Chemistry for Soil monitoring. In: Sparks, D.L. (Ed.), Advances in Agronomy, Vol. 132. Academic Press, pp. 139–159. https://doi.org/10.1016/bs.agron.2015.02.002.

Omondiagbe, O.P., Lilburne, L., Licorish, S.A., MacDonell, S.G., 2023. Soil texture prediction with automated deep convolutional neural networks and population-based learning. Geoderma 436, 116521. https://doi.org/10.1016/j.geoderma.2023.116521.

Pan, S.J., Yang, Q., 2010. A Survey on transfer Learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191.

Poggio, M., Brown, D.J., Bricklemyer, R.S., 2017. Comparison of Vis–NIR on in situ, intact core and dried, sieved soil to estimate clay content at field to regional scales. Eur. J. Soil Sci. 68 (4), 434–448. https://doi.org/10.1111/ejss.12434.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T., 2013. The spectrum-based learner: a new local approach for modeling soil vis–NIR spectra of complex datasets. Geoderma 195–196, 268–279. https://doi.org/10.1016/j.geoderma.2012.12.014.

Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158 (1), 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025.

Shenk, J.S., Westerhaus, M.O., Berzaghi, P., 1997. Investigation of a LOCAL Calibration Procedure for near Infrared Instruments. J. Near Infrared Spectrosc. 5 (4), 223–232. https://doi.org/10.1255/jnirs.115.

Soil Science Division Staff. (2017). Soil Survey Manual (K. S. C. Ditzler, Ed.; 1st ed.). U.S. Department of Agriculture, Government Printing Office.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Chapter Five—Visible and Near Infrared Spectroscopy in Soil Science. In: Sparks, D.L. (Ed.), Advances in Agronomy, Vol. 107. Academic Press, pp. 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7.

Tóth, G., Jones, A., Montanarella, L., 2013. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environ. Monit. Assess. 185 (9), 7409–7425. https://doi.org/10.1007/s10661-013-3109-3.

Tümsavaş, Z., Tekin, Y., Ulusoy, Y., Mouazen, A.M., 2019. Prediction and mapping of soil clay and sand contents using visible and near-infrared spectroscopy. Biosyst. Eng. 177, 90–100. https://doi.org/10.1016/j.biosystemseng.2018.06.008.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1), 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007.

Wang, D.-C., Zhang, G.-L., Zhao, M.-S., Pan, X.-Z., Zhao, Y.-G., Li, D.-C., Macmillan, B., 2015. Retrieval and Mapping of Soil Texture based on Land Surface Diurnal Temperature Range Data from MODIS. PLoS One 10 (6), e0129977. https://doi.org/10.1371/journal.pone.0129977.

Wang, Y., Borundia, A., Mouazen, A.M., 2025. Evaluating the economic and environmental effects of variable depth subsoiling in silage maize. Precis. Agric. 26 (6), 88. https://doi.org/10.1007/s11119-025-10284-8.

Wang, Y., Veltkamp, D. J., & Kowalski, B. R. (2002). Multivariate instrument standardization. ACS Publications; American Chemical Society. Doi: 10.1021/ac00023a016.

Wijewardane, N.K., Ge, Y., Morgan, C.L.S., 2016. Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. Geoderma 267, 92–101. https://doi.org/10.1016/j.geoderma.2015.12.014.