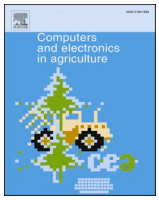
Computers and Electronics in Agriculture 239 (2025) 110920

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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



Original papers

Overfitting due to data leakage in soil sensor calibration: Examples from lab-based and *in-situ* soil NIR spectroscopy

José Correa [a], Hamed Tavakoli [a,*], Sebastian Vogel [a], Robin Gebbers [a,b]

[a] Department of Agromechatronics, Leibniz Institute for Agricultural Engineering and Bioeconomy e.V. (ATB), Max-Eyth-Allee 100, 14469 Potsdam, Germany
[b] Martin Luther University Halle-Wittenberg, Chair of Agricultural Business Operations, Karl-Freiherr-von-Fritsch-Straße 4, 06120 Halle, Germany

ARTICLE INFO

Keywords:
Proximal soil sensing
Principal component analysis (PCA)
Spatial interpolation
Machine learning
Pipeline

ABSTRACT

Sensor-based soil analysis methods, particularly optical spectroscopy, have gained as efficient alternatives to labor-intensive laboratory analyses for assessing soil properties. Especially *in-situ* measurements with mobile sensors streamline data collection, reducing both time and costs. This approach hinges on correlating sensor signals with laboratory-derived soil physico-chemical properties using mathematical calibration models. The models are trained and their parameters fine-tuned using a training dataset. It is best practice to evaluate the performance of calibration models by a test dataset, which is independent from the training dataset. However, certain commonly applied data preprocessing procedures can unintentionally introduce unwanted dependencies between the training and the test dataset. This is called data leakage. A model trained on these datasets will perform very well on both the training and test sets. However, it will show much poorer performance when tested on a truly independent dataset. Thus, the calibration model overfits via data leakage. In this study, we illustrate the consequences of data leakage by two common preprocessing procedures in soil sensing, namely principle component analysis (PCA) and spatial interpolation through ordinary kriging, on the prediction of soil properties by near infrared (NIR) spectroscopy. The NIR spectra were obtained in the laboratory and in the field. Laboratory measurements by standard wet-chemistry methods of soil pH value, total organic carbon (TOC) and total nitrogen (TN) content of 159 soil samples were used as target variables. Based on the results of this study, PCA and spatial interpolation led to data leakage when executed before data splitting. To avoid data leakage, we encourage researchers to carefully design leak-free data processing pipelines. These pipelines should encapsulate preprocessing methods, model fitting, and (if needed) spatial interpolation, ensuring that training and test sets are completely independent.

1. Introduction

In recent decades, soil sensing technologies have emerged as a compelling alternative to traditional soil analysis methods. These techniques offer several advantages, including rapid analysis, cost-effectiveness, and minimal sample preparation. Moreover, many soil sensing methods can be conducted directly on-site, enabling spatially high-resolution mapping of soil properties (Viscarra Rossel and Bouma, 2016). A sensor detects physical or chemical stimuli, such as heat, light, or electrochemical potential, and typically converts them into a variable voltage. This voltage is digitized, allowing computer processing. Successful sensing relies on a strong relationship between the measured stimuli and the property of interest, which is established through calibration. Calibration is an operation that, under specified conditions, in a first step, establishes a "relation" between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication (BIPM et al., 2012). This "relation" is typically represented by a mathematical function known as a calibration model or calibration function. Various methods, including statistics, chemometrics, machine learning, and geostatistics, are employed to establish these calibration models. While a simple linear regression modeling was a common approach in the past, spectral sensors and sensor combinations often require the use of more complex multivariate machine learning methods.

Since calibration is subjected to uncertainty, which we will briefly discuss below, the performance of calibration models must be evaluated.

* Corresponding author.
E-mail address: htavakoli@atb-potsdam.de (H. Tavakoli).

https://doi.org/10.1016/j.compag.2025.110920
Received 29 July 2024; Received in revised form 20 June 2025; Accepted 20 August 2025
Available online 11 September 2025

In machine learning, model evaluation by using independent training and test datasets is regarded as the standard (Ge et al., 2020; Hastie et al., 2009; Héberger et al., 2017; Joseph, 2022). The training dataset is used for model selection and model parameter tuning. The test dataset is used for assessing the model performance when applied to a new data domain. The resulting prediction error is also called test error or generalization error. In general, it is an indication of overfitting when models demonstrate good performance within the training set but exhibit poorer performance with unseen data in a test set (Ying, 2019). To obtain a reliable test error, it is important that the test samples are independent from the training set to avoid any influence of the test samples on the model training process and thus reduce the risk of overfitting on the training set (Kapoor and Narayanan, 2023; Zhu et al., 2023).

One cause of dependency in the data is the spatial and / or temporal dependency in the observations of a particular domain. However, dependencies can also be introduced by preprocessing of the data. Preprocessing methods (also called feature engineering), such as denoising, data transformation, and dimensionality reduction, are applied to spectral data to improve model calibration and prediction performance by removing physical interferences, fulfilling model assumptions, extracting and generating new features, or ensuring alignment (Mitchell, 2010; Rinnan et al., 2009; Xu et al., 2020; Zheng and Casari, 2018). Some of these methods work over the variables (features, columns) of a dataset. For example, spectra can be standardized by mean centering and/or variance scaling before multivariate modeling (Leone et al., 2012; Zhang and Hartemink, 2020; Baumann et al., 2021). Data standardization involves removing the mean (centering) and division by variance (scaling) of a feature to generate transformed data with zero mean and unit variance. However, this process introduces dependencies between observations, as each standardized value is influenced by the overall dataset statistics, such as the feature mean and variance. If standardization is applied before splitting the data into training and test sets, it can lead to unintended dependencies between the two sets (we provided an example of data leakage caused by centering and scaling in Supplementary File S1, using the LUCAS dataset). This phenomenon is known as “data leakage”, where information from the test set has leaked into the training process (Kapoor and Narayanan, 2023; Kaufman et al., 2011; Zhu et al., 2023). Kaufman et al. (2011) were the first who analyzed this problem systematically and coined the name “data leakage”. This term emerged after the phenomenon was previously referred as “leak from the future” in time series modeling, as described by Nisbet et al. (2018). Preventing data leakage is a good scientific practice, and most machine-learning practitioners are well aware of non-leaky workflow pipelines (Ge et al., 2020; Joseph, 2022; Yang et al., 2023). However, data leakage is a widespread problem in various fields, including medicine, genomics, and engineering, and any misused machine/deep learning technique can induce it (Hosseini et al., 2020; Kapoor and Narayanan, 2023; Kaufman et al., 2011; Rosenblatt et al., 2024; Yang et al., 2023; Zhu et al., 2023). Based on a literature survey in 17 fields of machine learning application Kapoor and Narayanan (2023) stated that data leakage has contributed to a “reproducibility crisis in ML-based data science”. According to Yang et al. (2023), nearly 30 % of 100,000 public Python notebooks show data leakage. This problem affected codes from all experience levels, including tutorials widely used for education. In a review of machine learning applications in environmental research by Zhu et al. (2023) data leakage was identified as a common pitfall in 148 highly influential papers. Data leakage can lead to models that do not generalize well and are prone to overfitting. Consequently, model performance metrics on the test set can be overly optimistic and may not reflect the true performance of the model on new and unseen data, potentially reducing confidence in their implementation (Kapoor and Narayanan, 2023; Muralidhar et al., 2021; Samala et al., 2021). One reason for problems in applying mathematical calibration methods in environmental sciences, including soil sensing, might be the confusing terminology. The methods are drawn from several disciplines such as chemometrics, multivariate statistics, geostatistics, pedometrics, machine learning, artificial intelligence, and data science. Consequently, there is no uniform body of theory and terminology.

In soil sensing, machine learning for sensor calibration is highly important because of the complexity and the huge uncertainties in the relationship between the original sensor signal and the target properties. Complexity arises through heterogeneity in the composition of the soil and its variability in space and time. Therefore, there is a multitude of possibly interfering factors, which is called matrix effect in chemometrics. Sensors showing a high selectivity are not strongly affected by the matrix. Selectivity is the extent to which a sensor can determine a particular analyte without interference from other components of the sample (Bănică, 2012). Among the common sensors used in proximal soil sensing, which include electrical resistivity meters, gamma-ray spectrometers, visible and near-infrared (Vis-NIR) spectrometers and cameras, as well as pH potentiometers (Gebbers, 2018), only the mobile potentiometric pH sensor (Adamchuk et al., 1999) shows high selectivity. On the contrary, the popular geoelectrical sensors show low selectivity since apparent electrical conductivity/resistivity, as measured by geoelectrical sensors, is affected by several soil properties like water content, texture, salinity, temperature, and bulk density (Corwin and Lesch, 2003). The laboratory methods, which are used to produce the reference data, try to deal with matrix effects by sample preprocessing (e.g., drying, sieving), homogenization, and extraction with buffered extractants over a longer extraction time. However, reproducibility within and between laboratories can sometimes be poor (Analytical Methods Committee, 2012). Even worse, the definition of certain target parameters, such as plant available phosphorous, is relatively vague and there are several different laboratory methods for assessing these parameters (Yli-Halla et al., 2016).

Among the abovementioned soil sensors, Vis-NIR spectroscopy combined with multivariate modeling techniques has gained a lot of interest because of its potential to predict a wide set of relevant soil properties (Ge et al., 2020; Hong et al., 2018; Stenberg et al., 2010). Soil organic matter (SOM), texture and nutrient content have been the primary targets (Stenberg et al., 2010). Soil Vis-NIR spectroscopy can be performed in the laboratory and in the field (Hong et al., 2018; Liu et al., 2017; Stenberg et al., 2010). However, Vis-NIR spectroscopy for soil analysis has two fundamental drawbacks: First, it is an indirect method. Many soil properties of interest show no absorption in the Vis-NIR range, and they are only assessed via cross-correlation with other soil properties that have absorption bands. These absorption bands are usually very wide and the bands are overlapping. Consequently, Vis-NIR spectra show only a few distinct features that can unequivocally be linked to soil properties of interest. This necessitates extra efforts for preprocessing spectra and the use of complex multivariate calibration techniques. Second, field spectral data is very susceptible to interference from external environmental factors, such as ambient light, temperature, soil moisture, soil structure, dust, stones, soil surface conditions, excessive crop residues, etc. (Hong et al., 2018; Stenberg et al., 2010). Although this results in mobile Vis-NIR spectroscopy being less accurate than predictions based on measurements of soil samples in the laboratory under standardized conditions, *in-situ* measurements still provide benefits due to their high spatial resolution and high measurement speed (Bönecke et al., 2021; Gebbers, 2018; Hong et al., 2018; Kodaira and Shibusawa, 2020; Vogel et al., 2022).

Soil Vis-NIR spectra can contain hundreds of highly correlated reflection bands, with many of them being overlapping or non-informative (Hidalgo et al., 2021; Hong et al., 2018; Liu et al., 2017). Therefore, one of the main steps in hyperspectral data processing is dimensionality reduction (DR) (Hidalgo et al., 2021; Hong et al., 2018; Liu et al., 2017). Many techniques have been developed for this purpose; however, principal component analysis (PCA) is one of the oldest and most widely used (Jolliffe and Cadima, 2016); especially in the context of optical spectroscopy (Xu et al., 2020). PCA projects high-dimensional

data to the direction of greatest variance by simultaneously satisfying the conditions of minimum error and maximum variance (Xu et al., 2020). In the case of spectral data, the original data matrix is converted to a group of new variables that are linear combinations of all wavelengths, namely principal components (PCs). A key step in PCA is the calculation of the covariance matrix, which contains all possible covariance values between each of the predictor variables. If PCA is done as DR strategy before data splitting, these covariances are involving the entire dataset including the test set, which may lead to data leakage.

In proximal soil sensing, a simple case of data leakage by preprocessing is the spatial alignment of geo-referenced data by interpolation. Spatial alignment is necessary if observations of predictor and dependent variables are not co-located, e.g., if sensor measurements and reference samples are not taken at the exact same positions. Spatial interpolation is used to estimate values at common positions. Since interpolation relies on information from neighboring observations, dependencies are introduced. Among spatial interpolation methods, ordinary kriging (OK) has traditionally been one of the most commonly used techniques in soil sensing (Gia Pham et al., 2019; Hengl, 2009; Schloeder et al., 2001). In recent years, machine learning and deep learning-based approaches for spatial interpolation have been proposed (Tziachris et al., 2020; Boumpoulis et al., 2023; Nwaila et al., 2024). However, a comprehensive review of these new methods is beyond the scope of this paper.

Data leakage is a well-studied topic in medicine, genomics, and engineering. However, no comprehensive survey or study on data leakage in soil science, particularly within soil proximal sensing, appears to exist in the current literature. This study aims to illustrate the consequences of data leakage on both laboratory- and field-based soil NIR spectroscopy. To investigate this, we conducted experiments with laboratory and field soil NIR data to explore how data leakage affects model training for predicting soil pH, total organic carbon (TOC), and total nitrogen (TN) content. As supplementary results and to generalize our illustration, we also used the Land Use/Land Cover Area Frame Survey (LUCAS) 2009 database, which includes laboratory-based Vis-NIR soil spectra and reference data from topsoil samples collected across different countries in the European Union (Nocita et al., 2014; Stevens et al., 2013) (Supplementary File S2). We also discussed the often-overlooked interaction between model complexity, data leakage, and spatial autocorrelation, factors that can significantly impact model reliability, particularly in the context of proximal soil sensing.. Finally, we provided some practical recommendations for evaluating data leakage risks in pipeline design.

2. Materials and methods

In this study, we calibrated regression models under various data leakage scenarios, utilizing spectral data collected from an agricultural study area in Germany (Supplementary Fig. S1). One dataset represents the case of field level calibration of spatially correlated data including *in-situ* and laboratory sensor measurements. It consists of field-collected NIR spectra captured by a mobile sensor platform (Tavakoli et al., 2022) and laboratory-measured NIR spectra from 159 soil samples. This dataset is named “Booßen”. We evaluated the models’ ability to predict key soil properties, including pH, total organic carbon (TOC), and total nitrogen (TN) content.

Additionally, we used a second dataset to illustrate the case of calibrating sensor measurements obtained in the laboratory under controlled conditions from wide-area data. It consists of data from the Land Use/Land Cover Area Frame Survey (LUCAS) 2009 database, which contains laboratory-based Vis-NIR soil spectra and reference data from topsoil samples collected across multiple European Union countries (Nocita et al., 2014; Stevens et al., 2013) (Section 2.4 and Supplementary File S1 and S2).

2.1. Study site and data collection

The study area of the Booßen dataset encompasses approximately 18.7 ha of agricultural land (52°23′38.69″N, 14°27′38.84″E) in eastern Brandenburg, Germany (Supplementary Fig. S1). Geologically, the parent material consists of end moraine glacial till, with Luvisol and Regosol as the dominant soil types. The soils are predominantly sandy, except for an area with higher clay content in the center of the field, probably representing deposits from an ancient stream (Schmidinger et al., 2024a). According to the German soil classification system KA5 (Eckelmann et al., 2005), the soil textures range mainly from slightly silty sand (Su2) to slightly loamy sand (Sl2). The elevation varies between 50 and 80 m.a.s.l., and the climate is characterized by 550 mm of annual rainfall and an average temperature of 9 °C (Schmidinger et al., 2024b). The sampling process took place along three parallel lines (12 m apart) running from the south-west to the northeast, capturing the main gradient of soil variability of the field (Schmidinger et al., 2024a). Along each transect line, 53 sample points were placed at 15-meter intervals, resulting in a total of 159 sampling points. Sampling was conducted between September 2nd and 4th, 2020. Sampling depth was 0 to 30 cm. The samples were analyzed in the laboratory for total nitrogen (TN, %), total organic carbon content (TOC, %) and soil pH (in $CaCl_2$). TN was determined according to DIN ISO 13878, TOC according to DIN ISO 10694, and soil pH (in $CaCl_2$) according to DIN ISO 10390. Descriptive statistics of the target variables are listed in Table 1.

Soil near infrared (NIR) diffuse reflectance was measured both *in-situ* in the field and on the 159 samples in the laboratory. In both cases, a NIR spectrometer (model C11118GA, Hamamatsu Photonics K. K., Shizuoka Pref., Japan) was used, covering the nominal spectral range of 860–2550 nm with an average resolution of 15 nm.

In December 2020, the laboratory measurements with the spectrometer were conducted after drying the samples, sieving them to less than 2 mm, and then filling them into a petri dish. Four halogen lamps at 45° illuminated the samples under controlled conditions inside a black box. The reflected light was guided to the spectrometer through an optical fiber. For each soil sample, the spectral data was collected with four replications, by mixing the soil in the petri dish and repeating the measurement. The spectrometer system was calibrated by measuring a certified reflection standard. The spectrum from the reference standard was used to obtain reflectance spectra from the samples by dividing the raw spectra by the reference spectrum. The average reflectance spectrum of each sample was then used for modeling. Because of the presence of noise, we had to remove some of the wavelengths at the two edges of the spectra. The final spectra had a range of 1,000 to 2,450 nm, which were interpolated to a resolution of 1 nm. The NIR spectra were preprocessed using the standard normal variate (SNV) transformation for each measurement. This carries no risk of data leakage as it involves row-wise transformations.

In-situ soil NIR measurements were conducted using a furrow-opening ‘shoe’ attached to a multi-sensor platform, as detailed in Tavakoli et al. (2024). These measurements were carried out during March and April 2020 (Supplementary Fig. S1). The shoe was pulled through the soil at an average speed of 2.5 km/h, allowing the measurement of subsurface soil reflectance at a depth of 10–15 cm. The measurements were carried out along parallel transects, about 25 m apart. The frequency of data collection was 1 Hz, which resulted in 18,906 measurement points. We applied the same edge removal procedure used for the laboratory data to the field spectra. The resulting spectra were in a range of 1,000 to 2,400 nm, which were also interpolated to a spectral resolution of 1 nm. We applied SNV normalization to each sensor measurement point individually.

2.2. Data analysis and data leakage scenarios

To investigate the impact of data leakage, laboratory- and field-based soil NIR spectroscopy data were calibrated to three target soil properties

Table 1
Descriptive statistics for the laboratory-measured target soil properties of the 159 soil samples taken at the test field.

Soil target property	Minimum	Q1	Median	Mean	SD	Q3	Maximum
TN (%)	0.04	0.07	0.09	0.11	0.06	0.14	0.28
TOC (%)	0.47	0.77	0.98	1.19	0.59	1.36	2.87
pH	5.03	5.74	6.15	6.32	0.74	7.14	7.54

Q1: first quartile; SD: standard deviation; Q3: third quartile.

across different leakage scenarios. PCA and kriging interpolation were applied either before or after data splitting, as shown in Table 2 and Figs. 1-7.

As a calibration modeling method we have chosen principal component regression (PCR) for this study to investigate the risk of data leakage associated with dimensionality reduction (DR) techniques like PCA. PCR integrates principal component analysis PCA (the DR step) followed by linear regression, making it ideal for simulating data leakage scenarios. PCR is less commonly used as compared to partial least squares regression (PLSR), which also relies on PCA and has a long history of successful applications in chemometrics (Geladi and Kowalski, 1986; Wold et al., 1993; Nørgaard et al., 2000; Wold et al., 2001) and is widely used in soil Vis-NIR spectroscopy (Brown et al., 2006; Mouazen et al., 2010). However, we restrict the investigation to PCR and do not include other calibration methods, such as PLSR, for the sake of clarity. Furthermore, PCR in our setting has proven to be significantly faster than PLSR during training, nearly three times faster, even when using a higher number of components, while resulting comparable predictive performance on our datasets. We selected 10-fold cross-validation to balance model performance and computational cost, given the characteristics of our dataset. Our data includes 159 reference samples and NIR spectra with 1,500 predictors measured across 18,000 geographical points, influencing our choice of data splitting strategy. *K*-fold cross-validation reduces variance in performance estimates, compared to a single hold-out set, especially with small datasets (Bishop, 2006). While leave-one-out cross-validation (LOO-CV) minimizes bias, it increases computational costs (Hastie et al., 2009), making it impractical for our field scenarios, which involves independent kriging for each predictor within each fold. For example, using a single hold-out set with 90 % of the data for training and 10 % for testing (15 samples in our case) can result in high variation in performance estimates due to the small size of the test set. *K*-fold cross-validation mitigates this variance by averaging performance over k different partitions, making the estimates less sensitive to data partitioning. However, a significant limitation of cross-validation is the substantial increase in computational cost associated with the k-fold training process. This is particularly true for our field scenarios, where we performed independent kriging for each of the 1,500 predictors within each fold. Therefore, to ensure a reliable evaluation of our models while keeping computational costs reasonable, we employed 10-fold cross-validation across all data leakage scenarios. With our dataset size of 159 samples, 10-fold cross-validation provides a good balance between accuracy and efficiency. In our dataset (Supplementary Fig. S2 and Table S1), as k increases, spatial interpolation execution time rises proportionally, despite fewer points per fold (Supplementary Table S1).The prediction error on the test set, RMSE, decreases with larger k values, with stability observed for $k > 8$ (Supplementary Fig. S2). However, each additional fold adds approximately 1 h to interpolation time. Using $k = 10$ strikes a balance between reducing variance and limiting computational demands, offering more reliable performance estimates than $k < 10$ while avoiding excessive costs associated with higher k values, especially during the kriging interpolation.

Table 2
Overview of laboratory- and field-based data leakage scenarios in soil NIR spectroscopy workflows using the Booßen dataset. See Figs. 1-7 for details.

Feature	Data leakage scenarios					
	L1 – Lab: PCA Before Split	**L2 – Lab: PCA in Pipeline (No Tuning)**	**L3 – Lab: PCA in Pipeline (Tuning)**	**F1 – Field: Full Interpolation**	**F2 – Field: Partial Independence**	**F3 – Field: Full Independence**
NIR Spectroscopy Setting	Laboratory-based	Laboratory-based	Laboratory-based	Field-based (mobile *in-situ*)	Field-based (mobile *in-situ*)	Field-based (mobile *in-situ*)
Cross-Validation Design	10-fold CV	10-fold CV	Nested CV (outer + inner)	Nested CV (outer + inner)	Nested CV (outer + inner)	Nested CV (outer + inner)
Spatial Interpolation Timing	—	—	—	Before all splitting	After outer split; before inner split	After all splitting
Spatial Interpolation Scope	—	—	—	Whole dataset	Per outer split	Per inner and outer fold
Train/Test Independence	CV folds: ×	CV folds: ✓	Outer: ✓ Inner: ✓	Outer: × Inner: ×	Outer: ✓ Inner: ×	Outer: ✓ Inner: ✓
PCA Application	Before CV split (leaky)	After CV split (in pipeline)	After inner split (in pipeline)	After outer and inner split (in PCR pipeline)	After outer and inner split (in PCR pipeline)	After outer and inner split (in PCR pipeline)
Hyperparameter Tuning	None	None	Grid search	Grid search on leaky data	Grid search with partial leakage	Grid search on fully independent folds
Components Used	15 fixed PCs	15 fixed PCs	Optimal q (tuned)	Optimal q (tuned)	Optimal q (tuned)	Optimal q (tuned)
Risk of Data Leakage[a]	Yes	No	No	Yes	Yes	No

[a]Note that that other upstream processes in the pipeline, which are not considered in this work, may also contribute to data leakage. These risks may arise during measurement due to duplicate readings, during sensor data processing, or because of spatial autocorrelation within the study area.
CV: Cross-validation.
×: Non-independent datasets.
✓: Independent datasets.
PCA: Principal Component Analysis.
PCR: Principal Component Regression
PCs: Principal Components.
q: Optimal number of principal components.

A

Scenario L1

1. Perform PCA on NIR data of the entire dataset before splitting the data.
2. Set the number of components to 15, which captures approximately 99.99% of the total variance of the NIR spectra.
3. Use the resulting 15 components as new predictor variables.
4. Perform a 10-fold cv by randomly shuffling the dataset (15 components and one soil target variable) and dividing it into 10 complementary folds of approximately 10% of the total data each.
5. (**Loop for model evaluation**) For fold k in the 10 folds:
   a) Leave fold k as the cv-test set.
   b) Use the remaining nine folds as the cv-training set.
   c) Fit a PCR model on the cv-training set using a linear regression with the 15 components as predictor variables to predict the soil target variable.
   d) Test the trained PCR model using the 15 components of the cv-test set as predictor variables.

B

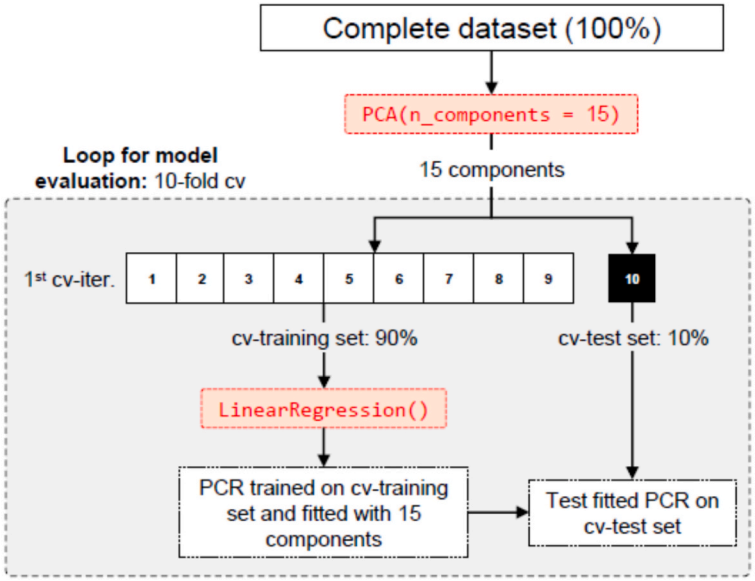


Fig. 1. Scenario L1: PCA before data splitting and without hyper-parameter tuning. **A**: Steps for model training and testing *via* a 10-fold cross-validation (cv). **B**: A diagram showing the first iteration of cross-validation loop, in which 10% of the data (fold 10, black square) was left as the cv-test set. The remaining 90% of the data (folds 1–9, white squares) was left as the cv-training set. PCA: principal component analysis. PCR: principal component regression.

2.2.1. Data leakage scenarios for laboratory-based soil NIR spectroscopy

2.2.1.1. Scenario L1: PCA before data splitting and without hyper-parameter tuning. Scenario L1 assesses the prediction performance of a model trained and tested under a risk of data leakage. For that, an individual PCA was applied on the reflectance values of the entire dataset. The number of components was heuristically set to 15. They explained more than 99.99 % of the variance of the NIR spectra (Supplementary Fig. S3). Then, using these 15 PCs as predictor variables, a linear regression was trained and tested via 10-fold cross-validation. Fig. 1 shows the different steps of this scenario in details.

2.2.1.2. Scenario L2: PCA pipelined into a PCR model without hyper-parameter tuning. In the L2 scenario, both cross-validation training and test sets were independent from each other to prevent data leakage. In each cv-iteration, a PCA was applied only on the NIR spectra of the cross-validation training set. As in scenario L1, the number of components was set to 15 and used as predictor variables for a linear

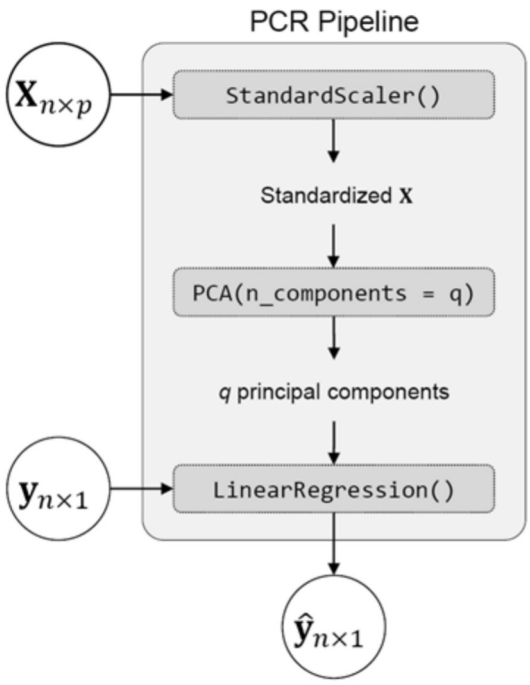


Fig. 2. Diagram of a pipeline for a principal component regression (PCR) model. The name of each function refers to the Python Scikit-learn library (version 0.23.1).

regression, which was validated on the cv-test set during each cv-iteration (Table 2). For this scenario, the PCA was encapsulated into a PCR pipeline as shown in Fig. 2. First, the spectra with n samples and p wavelength bands, $\mathbf{X}_{(n\times p)}$, are standardized using normal scores transformation. Mean and variance from standardizing the training set were stored for later application to the test set, ensuring consistent preprocessing (Pedregosa et al., 2011). The standardized $\mathbf{X}_{(n\times p)}$ was used as input for a principal component analysis (PCA) to obtain $q = 15$ principal components. After that, the q principal components and the soil target variable $\mathbf{y}_{(n\times 1)}$ of the same samples were used as predictor and target variables, respectively, for creating a simple linear regression. Finally, the predicted soil property, $\hat{\mathbf{y}}_{(n\times 1)}$, was obtained as output. In summary, a pipeline is a sequence of data preprocessing and modeling steps that can be treated as a single unit. Thus, during the whole cross-validation loop, 10 PCAs were fitted on the cv-training set. Fig. 3 shows the different steps of this scenario in details.

2.2.1.3. Scenario L3: PCA pipelined into a PCR model with hyper-parameter tuning. Similar to scenario L2, this scenario assesses the prediction performance of a model trained and tested under a no data leakage condition. However, unlike scenarios L1 and L2, the optimal number of components in this scenario was tuned *via* a nested cross-validation (hyper-parameter tuning). Fig. 4 provides a detailed overview of the steps in scenario L3, while Table 2 highlights the key differences compared to the previous scenarios.. A nested cross-validation has an inner-cv and an outer-cv. The inner-cv serves to carry out the hyper-parameter tuning, while performance of the trained and tuned model is evaluated within each iteration of the outer-cv. For that, the dataset was randomly divided into 10 folds. During each outer-iteration, one fold was left as an outer-test set and the remaining nine as an outer-training set (Fig. 4). The optimal number of components was determined by fitting PCR models with 1 to 60 components within the inner-training set through a nine-fold inner cross-validation. Then the trained and tuned model was tested on the outer-test set. In this way, the test and training sets within each iteration of the cross-validation were independent from each other, at least in terms of PCA.

2.2.2. Data leakage scenarios for field-based (in-situ) soil NIR spectroscopy

The field NIR data was collected with high spatial resolution using a mobile sensor platform (Tavakoli et al., 2022). Since positions of the reference samples and the NIR measurements were not always exactly co-located, spatial interpolation of the NIR data to the reference sampling points was necessary (Supplementary Fig. S1). Therefore, we performed ordinary block kriging using the 'gstat' library in R (Pebesma, 2004) with a block size of 10×10 m^2. To expedite the interpolation process, we considered only the 500 measurement points that were spatially closest to each soil sampling point. Due to the large number of variables (ca. 1,400 wavelenghts), a variogram model (for kriging) was automatically fitted for each predictor variable using the 'automap' library of R (Hiemstra et al., 2009). The performance of kriging interpolation using 10-fold cross-validation for each field scenario is summarized in Supplementary Fig. S4.

To investigate the potential for data leakage in *in-situ* soil NIR spectroscopy, we designed three scenarios (shown in detail in Figs. 5-7 and summarized in Table 2) for interpolation either before or after data splitting as outlined in the next paragraphs. For these three scenarios, a pipelined PCR (Fig. 2) was trained using the interpolated NIR spectra as the predictor variables and the soil property of interest as the target variable. As the scenario L3, the optimal number of components was determined by fitting PCR models with 1 to 60 components within the inner-training set through a nine-fold inner cross-validation.

2.2.2.1. Scenario F1: Interpolation of the complete dataset as a whole. In this scenario, the *in-situ* NIR data was interpolated to soil sampling points using the entire dataset before data splitting. After the interpolation, a pipelined PCR model was trained and tested on the data *via* a nested cross-validation (Fig. 2). Fig. 5 gives details about the different steps of this scenario.

2.2.2.2. Scenario F2: Independent interpolation of the training and test sets. In this scenario, in each iteration of the outer cross-validation, the outer-training and outer-test sets were independently interpolated (Fig. 6). This means that two separate interpolations were applied to the reflectance values of the training and test sets, ensuring that they remained independent. Note that the hyper-parameter tuning is done on interdependent folds, which are connected by a common interpolation run. Fig. 6 shows the different steps of this scenario in detail.

2.2.2.3. Scenario F3: Independent interpolation within each fold of cross validation. To prevent data leakage, separate kriging interpolations were applied on the reflectance values of each of the 10 outer cross-validation folds (Fig. 7). It is important to note that unlike scenario F2, the hyper-parameter tuning was performed on completely disconnected folds that were not connected by a common interpolation run (Table 2).

2.2.3. Spatial autocorrelation

Spatial autocorrelation (SAC) arises when measurements from nearby locations exhibit higher similarity or lower dissimilarity compared to those from distant locations (Beale et al., 2010; Hurlbert, 1984; Legendre, 1993; Tobler, 1970). SAC can lead to overfitting due to data leakage (Karasiak et al., 2022) and artificially inflating model performance (Kattenborn et al., 2022; Ploton et al., 2020). Hence, assessing SAC is vital for accurate modeling and model validity. To explore the degree of spatial autocorrelation of our results, the Moran's I autocorrelation coefficient (see Dormann et al. (2007)) was calculated both for the residuals and target variables. The coefficient varies in the range of -1 to $+1$. Positive and negative values indicate positive and negative spatial autocorrelations, respectively. Zero value means no spatial autocorrelation (null hypothesis). The closer the values to zero, the less the spatial autocorrelation. The analysis was conducted using the R package "ape" (Paradis and Schliep, 2018). For this purpose, the

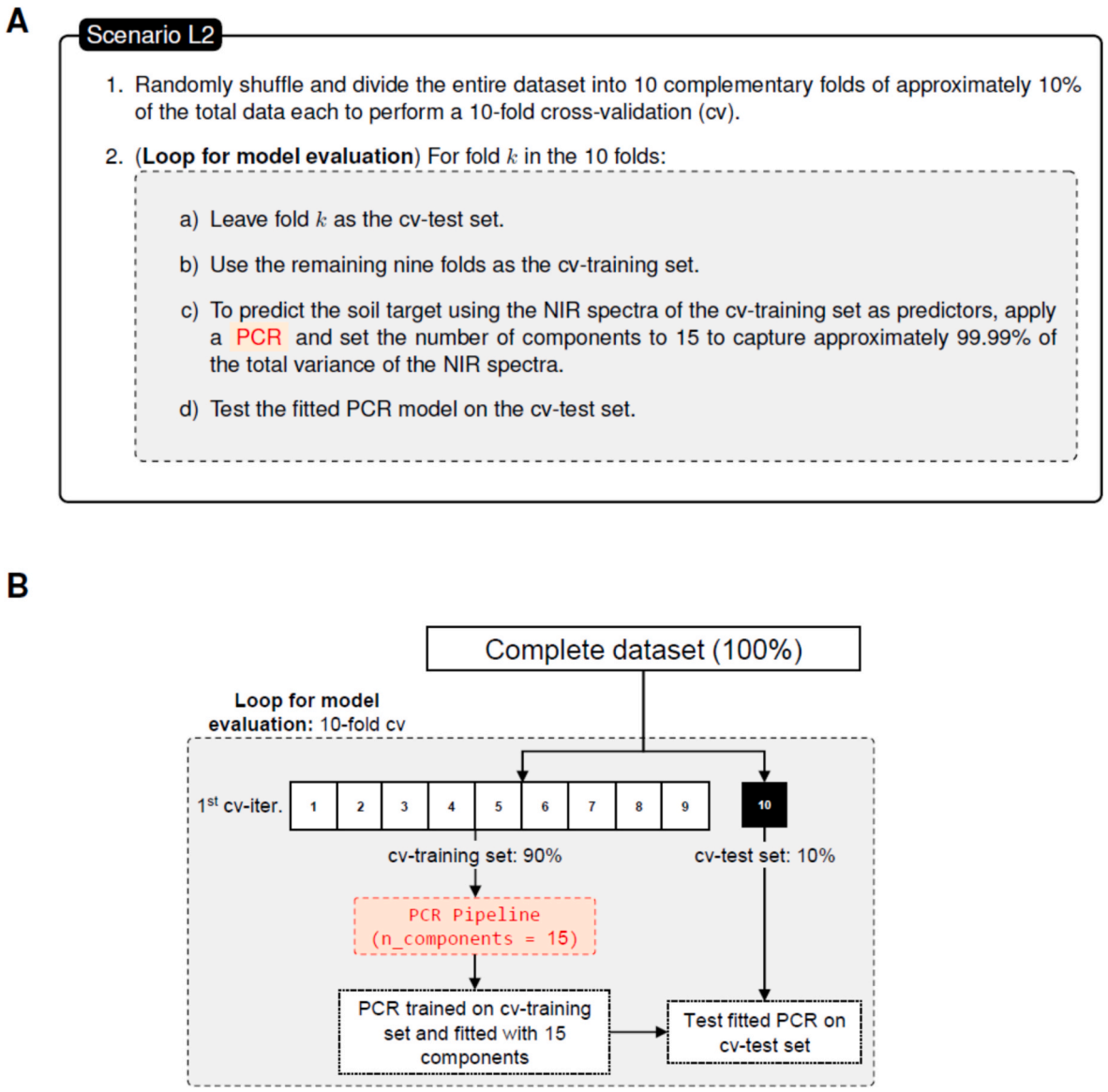


Fig. 3. Scenario L2: PCA pipelined into a PCR model without hyper-parameter tuning. **A**: Steps for model training and testing *via* a 10-fold cross-validation (cv). **B**: A diagram showing the first cv iteration in which 10% of the data was left as the cv-test set (fold 10, black square). The remaining 90% of the data was left as the cv-training set (folds 1–9, white squares). The model training was based on a principal component analysis (PCA) pipelined into a principal component regression (PCR).

elements of the matrix of spatial weights were obtained according to the negative exponential function (more details can be found in Chen (2013) and Chen (2016)). Additionally, model residuals were plotted on a map to check possible patterns of spatial autocorrelation.

2.2.4. Example of PCA-mediated data leakage using the LUCAS dataset (Supplementary File S2)

To exemplify the occurrence of overfitting arising from data leakage during soil sensor calibration within a larger dataset, we employed the comprehensive LUCAS dataset (Nocita et al., 2014; Stevens et al., 2013). Three distinct scenarios were examined, each involving the application of principal component analysis (PCA) on Vis-NIR spectra:

1. Scenario 1: Leaked PCA – PCA performed on the entire dataset before data splitting, without hyper-parameter tuning for the optimal number of components.
2. Scenario 2: Leaked PCA with Tuning – PCA was performed on the entire dataset before data splitting, with the optimal number of components determined through cross-validation on the entire dataset.
3. Scenario 3: Non-Leaked PCA – PCA was integrated into a PCR model using a pipeline.

In all scenarios, principal components were used as predictors within a linear regression model. Subsequently, the trained models were evaluated on a completely independent dataset. Detailed methodological descriptions for each step are provided in Supplementary File S2.

2.2.5. Model performance metrics

The prediction performance of the cross-validated models were assessed using the ratio of performance to interquartile (RPIQ), the root mean squared error (RMSE) and the coefficient of determination (R^2). Those performance metrics are calculated as follows:

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

A

Scenario L3

1. Randomly shuffle and divide the entire dataset into 10 complementary folds of approximately 10% of the total data each to perform a 10-fold cross-validation (cv).
2. (**Outer loop for model evaluation**) For fold k in the 10 folds:
   a) Set fold k as the outer-test set.
   b) Set the remaining nine folds as the outer-training set.
   c) Tune the optimal number of components within the outer-training set *via* a nine-fold cv (the inner-cv).
   d) (**Inner loop for hyper-parameter tuning**) For fold l in outer-training set:
      i. Set fold l as the inner-test set.
      ii. Leave the eight remaining folds of the outer-training set as the inner-training set.
      iii. Train PCRs with one to q components on the inner-training set and test on the inner-test set.
   e) Select the optimal number of components as the one that presents the lowest average MSE value among all the inner-iterations.
   f) Train a PCR with the tuned number of components on the outer-training set and test on the outer-test set.

B

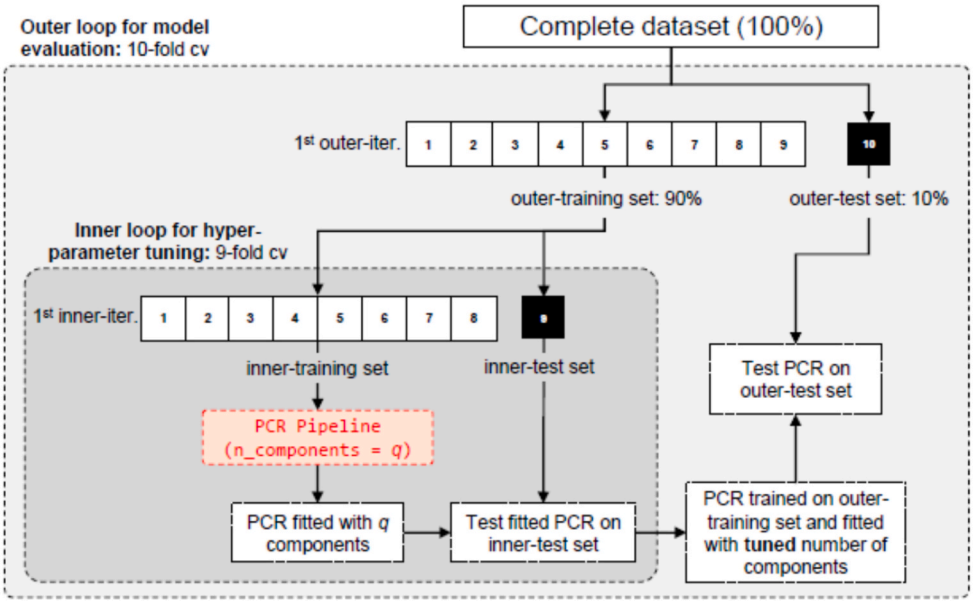


Fig. 4. Scenario L3: PCA pipelined into a PCR model with hyper-parameter tuning. **A**: Steps for model training and testing *via* a nested cross-validation (cv). **B**: A diagram showing the first outer-cv iteration in which 10% of the data was left as the outer-test set (fold 10, black square). The remaining 90% of the data was left as the outer-training set (folds 1–9, white squares). In the first iteration of the inner loop, the outer training set was further divided into two subsets: the inner-training set and the inner-test set. The model training was based on a principal component analysis (PCA) pipelined into a principal component regression (PCR).

A

Scenario F1

1. Interpolate the NIR spectra to soil sample points using the entire dataset before data splitting.
2. Shuffle the interpolated spectra and soil target variable, and divide them into 10 complementary folds, each containing approximately 10% of the total data.
3. (**Outer loop for model evaluation**) For fold k in the 10 folds:
   a) Set fold k as the outer-test set.
   b) Use the remaining nine folds as the outer-training set.
   c) Tune the optimal number of components within the outer-training set *via* a nine-fold cv (the inner-cv).
   d) (**Inner loop for hyper-parameter tuning**) For fold l in outer-training set:
      i. Set fold l as the inner-test set.
      ii. Leave the eight remaining folds of the outer-training set as the inner-training set.
      iii. Train PCRs with one to q components on the inner-training set and test on the inner-test set.
   e) Select the optimal number of components as the one that presented the lowest average MSE value among all the inner-iterations.
   f) Train a PCR with the tuned number of components on the outer-training set and test it on the outer-test set.

B

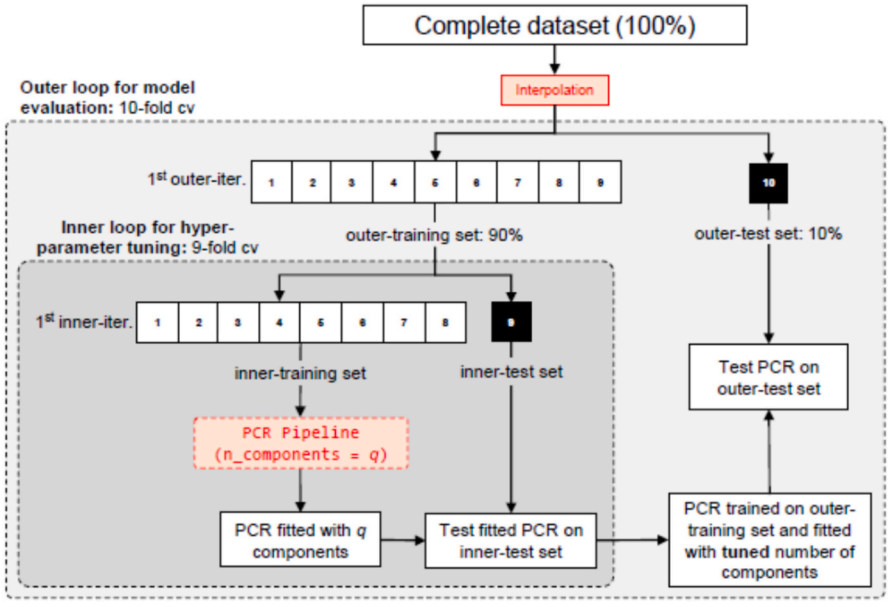


Fig. 5. Scenario F1: Interpolation of the complete dataset as a whole. **A**: Steps for model training and testing *via* a nested cross-validation (cv). **B**: A diagram showing the first outer-cv iteration in which 10% of the data (fold 10, black square) was left as the outer-test set. The remaining 90% of the data (folds 1–9, white squares) was left as the outer-training set. In the first iteration of the inner loop, the outer training set was further divided into two subsets: the inner-training set and the inner-test set. The model training was based on a principal component analysis (PCA) pipelined into a principal component regression (PCR).

A

Scenario F2

1. Randomly shuffle the complete dataset and divide it into 10 complementary folds, each containing approximately 10% of the total data, to perform a 10-fold cv (outer-cv).
2. (**Outer loop for model evaluation**) For fold k in the 10 folds:
   a) Set fold k as the outer-test set.
   b) Use the remaining nine folds as the outer-training set.
   c) Interpolate NIR spectra to soil sample points using the outer-training set as a whole.
   d) Interpolate NIR spectra to soil sample points using the data of the outer-test set.
   e) Tune the optimal number of components within the outer-training set *via* a nine-fold cv (inner-cv).
   f) (**Inner loop for hyper-parameter tuning**) For fold l in outer-training set:
      i. Set fold l as the inner-test set.
      ii. Leave the eight remaining folds of the outer-training set as the inner-training set.
      iii. Train PCRs with one to q components on the inner-training set and test on the inner-test set.
   g) Select the optimal number of components as the one that presented the lowest average MSE value among all the inner-iterations.
   h) Train a PCR with the tuned number of components on the outer-training set and test it on the outer-test set.

B

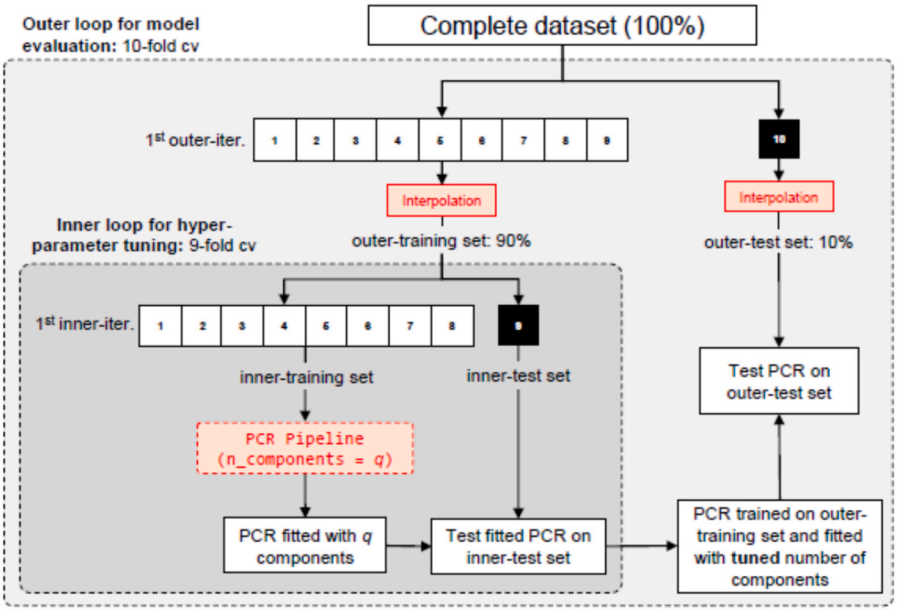


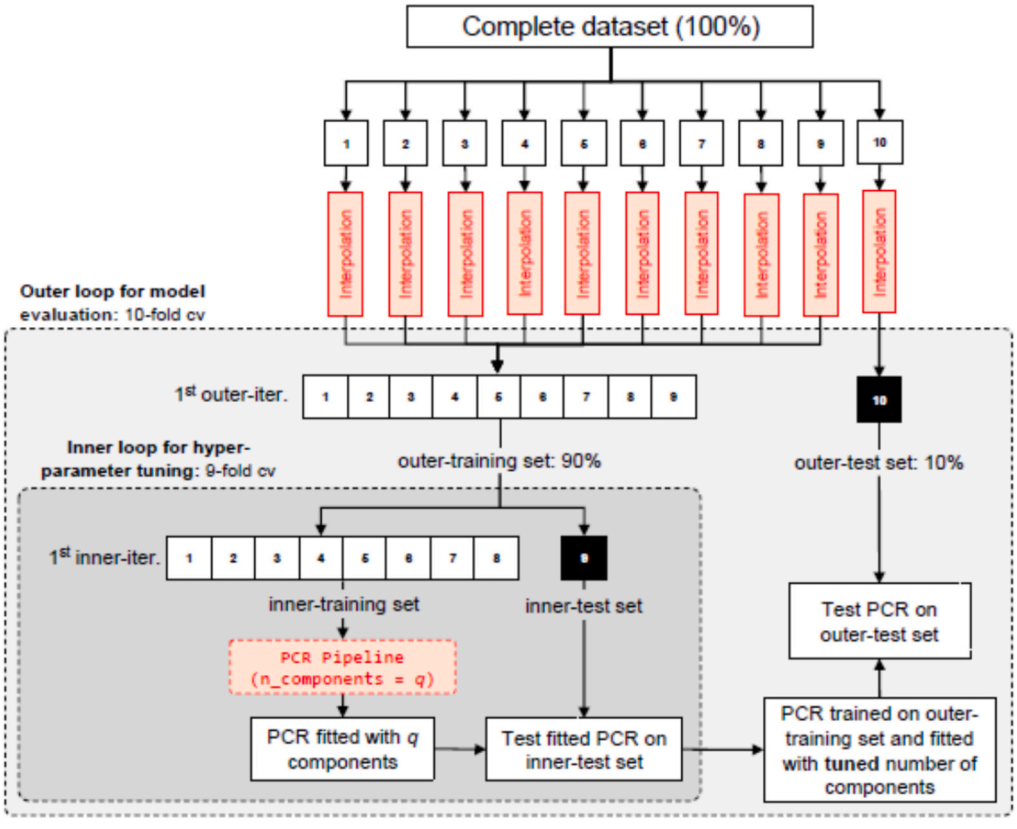
Fig. 6. Scenario F2: Independent interpolation for the training and test sets. **A**: Steps for model training and testing *via* a nested cross-validation (cv). **B**: A diagram showing the first outer-cv iteration in which 10% of the data (fold 10, black square) was left as the outer-test set. The remaining 90% of the data (folds 1–9, white squares) was left as the outer-training set. In the first iteration of the inner loop, the outer training set was further divided into two subsets: the inner-training set and the inner-test set. The model training was based on a principal component analysis (PCA) pipelined into a principal component regression (PCR).

A

Scenario F3

1. Shuffle and divide the complete dataset into 10 complementary folds, each containing about 10% of the total data, for performing a 10-fold cross-validation (the outer-cv) before spectra interpolation.
2. Interpolate the sensor NIR spectra to soil sample points within each fold.
3. (**Outer loop for model evaluation**) For fold k in the 10 folds:
   a) Set fold k as the outer-test set.
   b) Use the remaining nine folds as the outer-training set.
   c) Tune the optimal number of components within the outer-training set *via* a nine-fold cv (inner-cv).
   d) (**Inner loop for hyper-parameter tuning**) For fold l in outer-training set:
      i. Set fold l as the inner-test set.
      ii. Leave the eight remaining folds of the outer-training set as the inner-training set.
      iii. Train PCRs with one to q components on the inner-training set and test on the inner-test set.
   e) Select the optimal number of components as the one that presented the lowest average MSE value among all the inner-iterations.
   f) Train a PCR with the tuned number of components on the outer-training set and test it on the outer-test set.

B



(caption on next page)

Fig. 7. Scenario F3: Independent interpolation within each fold of cross-validation. **A**: Steps for model training and testing *via* a nested cross-validation (cv). **B**: A diagram showing the first outer-cv iteration in which 10% of the data (fold 10, black square) was left as the outer-test set. The remaining 90% of the data (folds 1–9, white squares) was left as the outer-training set. In the first iteration of the inner loop, the outer training set was further divided into two subsets: the inner-training set and the inner-test set. The model training was based on a principal component analysis (PCA) pipelined into a principal component regression (PCR).

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

$$RMSE = \sqrt{MSE} \quad (3)$$

$$RPIQ = \frac{IQR}{RMSE} \quad (4)$$

where n is the number of testing samples; y_i is the observed value for the target variable y of the soil sample i; $\hat{y}_i$ is the predicted values of the sample i; $\bar{y}$ is the mean of the observed values of the samples ($\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$); IQR is the interquartile range of the observed values.

2.2.6. Software and hardware platform

Data management and plotting were done in R (R Core Team, 2019) using packages from 'tidyverse' (Wickham et al., 2019). Models were calibrated using the Scikit-learn (version 0.23.1; (Pedregosa et al., 2011)) library of Python (v3.7.10; https://www.python.org). Model training and data analyses were done on a single machine with the following specifications: Intel Core i7-10850H, 2.70 GHz, 6 cores, 12 logical processors; 16 GB RAM; Windows 10.

3. Results and discussion

3.1. Results for lab-based soil NIR spectroscopy

The results obtained for predicting the soil properties, total nitrogen (TN), total organic carbon (TOC) and pH by applying the three laboratory scenarios are presented in Table 3. In addition, Fig. 8 illustrates scatterplots of measured soil properties versus predicted ones for each model on the test set. To summarize the results, the predictions on the 10 training and test sets of the cross-validation were pooled to calculate performance metrics and graph the scatterplots. In all cases across the laboratory scenarios, the models performed better in the training set than in the test set (on average 0.90 and 0.50 in terms of R^2, respectively). This large difference indicates overfitting in certain scenarios (Table 3).

Among all the scenarios, in general, the best performances on the test set were attributed to scenario L3 (mean R^2 of 0.94, 0.88 and 0.85 for TN, TOC and pH, respectively) followed by scenario L1 (mean R^2 of 0.91, 0.88 and 0.80 for TN, TOC and pH, respectively) (Table 3). In contrast, scenario L2 exhibited the poorest performance across all target soil properties, with R^2 values of 0.91 and −0.24 on training and test sets, respectively. This discrepancy is further highlighted by the considerably higher average RMSE on the test set for L2, which was two to five times greater than that of L1 and L3 (Table 3). A negative R^2 indicates that the difference between the true values and the predictions (residual sum of squares) is greater than the difference between the true values and the mean (total sum of squares; (see Eq. (1). In the present study, scenario L2 showed a highly negative R^2 value and high RMSE values in the test set, indicating that this model is overfitted and has poor predictive performance (Table 3 and Fig. 8). RPIQ was consistent across the scenarios for different soil target variables, with scenario L2 having the highest RPIQ in the training set and the lowest RPIQ in the test set, indicating overfitting. In contrast, scenarios L1 and L3 consistently performed well on both training and testing sets.

In scenario L1, PCA was done before data splitting and no information about the target variable was considered to tune the number of components (Table 2). This introduced a dependency between the training and test sets. Under this scenario, the good performance observed on the test set is due to data leakage. This becomes evident when comparing the results with those of scenario L2 (Table 3). The only difference between these two scenarios is that L1 used a leaky pipeline while L2 was based on a non-leaky pipeline (Table 2). Both scenarios had the same number of untuned components. In both scenarios, linear regression was trained and tested using 15 PCs as predictor variables. However, in scenario L2, the cv-training and cv-test sets at each cross-validation iteration were not subjected to a common PCA to avoid any dependency between them. Despite this, scenario L2 showed the worst performance. This means that the determination of the optimal number of components is a key to achieve a good performance and that their heuristic determination fails to generalize well in unseen data. This observation is consistent with the findings obtained through the analysis of the larger, more general LUCAS dataset. As elaborated in Supplementary File S2, applying PCA to the entire dataset prior to data splitting, as exemplified in LUCAS scenarios 1 and 2, leads to the creation of non-independent training and test sets. This consequently results in overfitted regression models characterized by artificially high test

Table 3
Mean model performance on cross-validation test and training sets for each soil target variable under the three laboratory scenarios (mean ± std).

Soil target variable	Scenario*	Set	RPIQ**	RMSE	R^b
TN (%)	L1	Training	4.71 ± 0.45	0.01 ± 0.00	0.94 ± 0.00
		Test	3.93 ± 1.89	0.02 ± 0.00	0.91 ± 0.04
	L2	Training	4.84 ± 0.53	0.01 ± 0.00	0.95 ± 0.00
		Test	0.99 ± 0.49	0.07 ± 0.02	−0.34 ± 0.70
	L3	Training	7.01 ± 0.89	0.01 ± 0.00	0.97 ± 0.00
		Test	4.54 ± 2.00	0.01 ± 0.00	0.94 ± 0.02
TOC (%)	L1	Training	3.94 ± 0.30	0.16 ± 0.01	0.93 ± 0.01
		Test	3.60 ± 1.70	0.18 ± 0.05	0.88 ± 0.08
	L2	Training	4.00 ± 0.34	0.15 ± 0.01	0.93 ± 0.01
		Test	0.98 ± 0.52	0.66 ± 0.20	−0.36 ± 0.70
	L3	Training	4.73 ± 0.38	0.13 ± 0.01	0.95 ± 0.01
		Test	3.45 ± 1.66	0.18 ± 0.04	0.88 ± 0.07
pH	L1	Training	5.03 ± 0.14	0.28 ± 0.00	0.86 ± 0.01
		Test	3.77 ± 1.77	0.31 ± 0.05	0.80 ± 0.07
	L2	Training	5.07 ± 0.18	0.28 ± 0.01	0.86 ± 0.01
		Test	1.69 ± 0.54	0.71 ± 0.27	−0.01 ± 0.81
	L3	Training	6.29 ± 0.36	0.22 ± 0.01	0.91 ± 0.01
		Test	4.27 ± 1.90	0.28 ± 0.05	0.85 ± 0.06

* In the L3 scenario, the mean performance is calculated based on the outer-training set for the training set and the outer-test set for the test set.
**RPIQ: ratio of performance to interquartile range; RMSE: root mean squared error; R^2: the coefficient of determination. L1: PCA before data splitting and without hyper-parameter tuning; L2: PCA pipelined into a PCR model without hyper-parameter tuning; L3: PCA pipelined into a PCR model with hyper-parameter tuning.

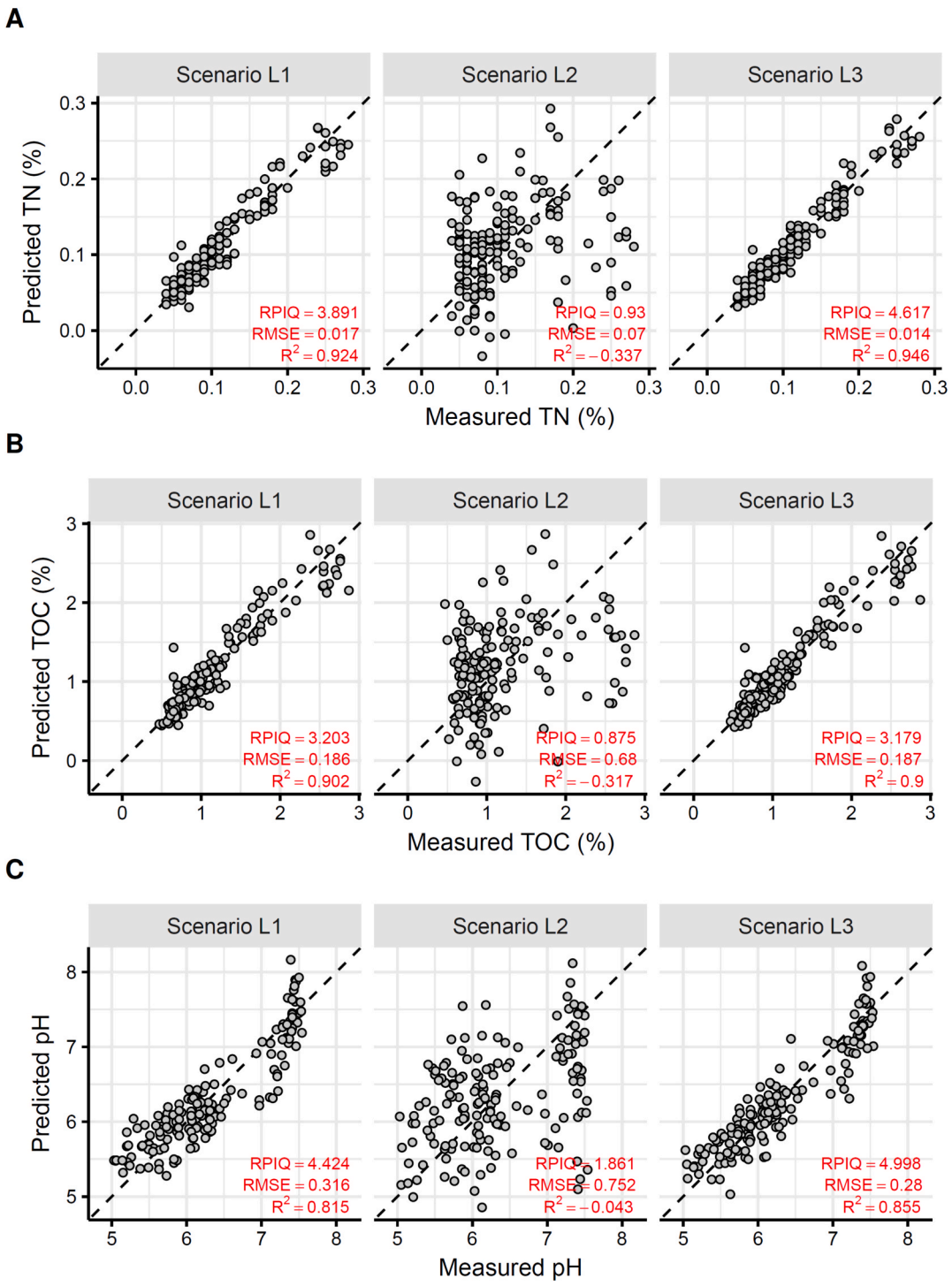


Fig. 8. Prediction performance of trained models on the test set under the three laboratory scenarios for A) soil total nitrogen (TN), B) soil total organic carbon (TOC), and C) soil pH (pH). **L1**: PCA before data splitting and without hyper-parameter tuning; **L2**: PCA pipelined into a PCR model without hyper-parameter tuning; **L3**: PCA pipelined into a PCR model with hyper-parameter tuning. RPIQ: ratio of performance to interquartile range; RMSE: root mean squared error; R^2: the coefficient of determination. The black and dashed line is the 1:1 line.

performance while demonstrating poor generalization capabilities when confronted with completely independent data. Moreover, the process of hyper-parameter tuning, such as the selection of the optimal number of components, is also compromised in LUCAS scenario 2, ultimately leading to models that exhibit poor performance when evaluated on unseen datasets.

The performance of the PCR model on the test set in scenario L1 was inferior to that observed in scenario L3. This may be attributed to the fixed number of components (15), which may have been insufficient to overfit the data to the fullest extent possible. It is well known that increasing the number of components can lead to a higher degree of overfitting. Hence, while the use of a larger number of components is likely to improve the performance of the model, it is also expected to result in a greater degree of overfitting, particularly in scenario L1 compared to scenario L3.

Within scenario L2, PCA was done after data splitting but no

information about the target variable was considered for tuning the number of components (Table 2). This led to independency between the training and test sets, however resulting in a model not being general enough to predict new data. According to Table 3 and Fig. 8, models trained under this scenario suggest overfitting. While they performed well on the training data, their performance on the unseen test data was significantly worse, particularly compared to scenarios L1 and L3. In both L2 and L3, no hyperparameter tuning was performed, and both models shared the same number of components (Table 2). Comparing these scenarios highlights the impact of data leakage, as seen in the higher test-set performance of L1 (Table 3), where data leakage was present, compared to L2, which follows a no data leakage approach.

In scenario L3, PCA was pipelined into a PCR model (Fig. 2). This way, the test and training sets within each iteration of the cross-validation were independent of each other, at least in terms of standardization and PCA (Figs. 3 and 4). As implemented in scikit-learn, if correctly pipelined, standardization first computes the mean and standard deviation from the training set and stores them for later use in centering and scaling the test set (Pedregosa et al., 2011). Since the tuning of the optimal number of components was based on a PCR, the information on the target variable was considered. This is an additional reason why the performance of the model under this scenario was better as compared to scenario L1. Since the training and hyper-parameter tuning steps were done in independent data sets, the tuned model of scenario L3 was much better than the other scenarios in predicting the target variable in the outer-test set (Table 3 and Fig. 8). Thus, optimal hyper-parameter tuning should be done using independent data sets to prevent data leakage. Additionally, when data leakage is not an issue, hyper-parameter tuning is preferred over heuristic methods for determining the number of components. Similar results were observed using the LUCAS dataset (Supplementary File S2). In the non-leaked LUCAS scenario 3, where PCA was integrated into a PCR model within a pipeline, independence between training and test sets was maintained during each cross-validation iteration. By performing model training and hyper-parameter tuning on independent datasets, scenario 3 demonstrated superior predictive performance on unseen data compared to the scenarios involving data leakage.

The performance of L1 is attributed to data leakage, as evidenced by its comparison with L2. However, despite suffering from overfitting due to leakage, L1 still performed slightly worse than L3. This suggests that L3′s superior performance is not solely due to its lack of data leakage but also because it incorporates appropriate hyperparameter tuning, which enhances model generalization. Furthermore, the lower performance of L1 may be attributed to the small number of PCA components selected for this scenario, which restricted the model's ability to further overfit and, at least artificially, outperform L3.

Since L2 and L3 are both no leakage scenarios (Table 2), L2′s poor test-set performance is due to its selected number of components lacking generalization for accurate predictions. Hyperparameter tuning is essential to prevent overfitting, as evidenced by L3′s improved performance (Table 3). Therefore, optimal hyperparameter tuning should always be conducted on independent datasets to prevent data leakage and ensure that the tuned hyperparameters are general enough to predict unseen data, avoiding overfitting.

Results of the Booßen laboratory scenarios are consistent with those obtained using the LUCAS dataset (Supplementary File S2): Data leakage occurs during the calibration of soil Vis-NIR spectra to target soil properties using principle component analysis (PCA) when the trained model performs better on datasets that share a common PCA with the training set. This PCA-mediated data leakage leads to over-fitted models that perform poorly on unseen and independent datasets. Data leakage also affects hyper-parameter tuning.

When comparing our study with other laboratory-based studies which utilized the LUCAS dataset, such as Tavakoli et al. (2023) and Zhong et al. (2021) or those found in the Supplementary File S2, our model shows better performance in terms of R^2 for TN under the L3 scenario (0.94 versus 0.88–0.93). However, our results suggest that the models developed by Tavakoli et al. (2023), in which similar pipeline has been used, have a superior overall fit to the data for TOC and pH compared to our models ($R^2 \sim 0.88$ versus 0.95 and 0.85 versus 0.94, respectively). In terms of R^2 values, the models developed by Tavakoli et al. (2023) can explain a larger proportion of the variability observed in the target variables in the LUCAS dataset. Nonetheless, it is important to bear in mind that comparing R^2 values across different datasets has limitations since it strongly depends on the statistical distribution of the values (Alexander et al., 2015), and factors such as the sample size, geographic distribution, and measurement techniques can influence the model's performance.

3.2. Results from calibrating field-based (in-situ) soil NIR spectroscopy

In this research we did not the address to what extent the parameterization of kriging may influence model's performance and contribute to data leakage. Settings such as the number of nearest neighbors used for interpolating (e.g., the 'nmax' argument in the 'gstat::krige' function in R) could affect results and merit further investigation in a dedicated study.

Table 4 presents the results of predicting the soil properties for the three field scenarios on both outer-test and outer-training sets. Moreover, scatterplots of measured soil properties versus predicted ones for each model on the outer-test set are shown in Fig. 9. For presenting the

Table 4
Mean model performance on outer-test and outer-training sets for each soil target variable under the three field scenarios (mean ± std).

Soil target variable	Scenario	Set	RPIQ*	RMSE	R^b
TN (%)	F1	Training	6.21 ± 0.47	0.01 ± 0.00	0.97 ± 0.00
		Test	4.16 ± 1.64	0.02 ± 0.00	0.93 ± 0.02
	F2	Training	6.09 ± 0.54	0.01 ± 0.00	0.97 ± 0.01
		Test	2.83 ± 1.20	0.02 ± 0.01	0.83 ± 0.10
	F3	Training	4.57 ± 0.37	0.01 ± 0.00	0.94 ± 0.00
		Test	3.89 ± 2.31	0.02 ± 0.00	0.91 ± 0.04
TOC (%)	F1	Training	5.07 ± 0.57	0.12 ± 0.01	0.96 ± 0.01
		Test	3.58 ± 1.60	0.17 ± 0.05	0.90 ± 0.05
	F2	Training	5.72 ± 1.04	0.11 ± 0.02	0.97 ± 0.01
		Test	2.56 ± 1.39	0.25 ± 0.07	0.78 ± 0.13
	F3	Training	3.79 ± 0.52	0.16 ± 0.02	0.92 ± 0.01
		Test	3.27 ± 1.96	0.20 ± 0.09	0.87 ± 0.07
pH	F1	Training	5.59 ± 0.95	0.26 ± 0.04	0.88 ± 0.04
		Test	3.01 ± 0.96	0.38 ± 0.07	0.70 ± 0.14
	F2	Training	6.57 ± 2.22	0.23 ± 0.05	0.90 ± 0.04
		Test	2.2 ± 0.51	0.52 ± 0.15	0.46 ± 0.26
	F3	Training	4.69 ± 1.11	0.31 ± 0.07	0.81 ± 0.09
		Test	2.4 ± 0.59	0.48 ± 0.12	0.55 ± 0.19

* RPIQ: ratio of performance to interquartile range; RMSE: root mean squared error; R^2: the coefficient of determination. F1: Interpolation on the complete dataset as a whole; F2: Independent interpolation for the training and test sets; F3: Independent interpolation within each fold of cross-validation.

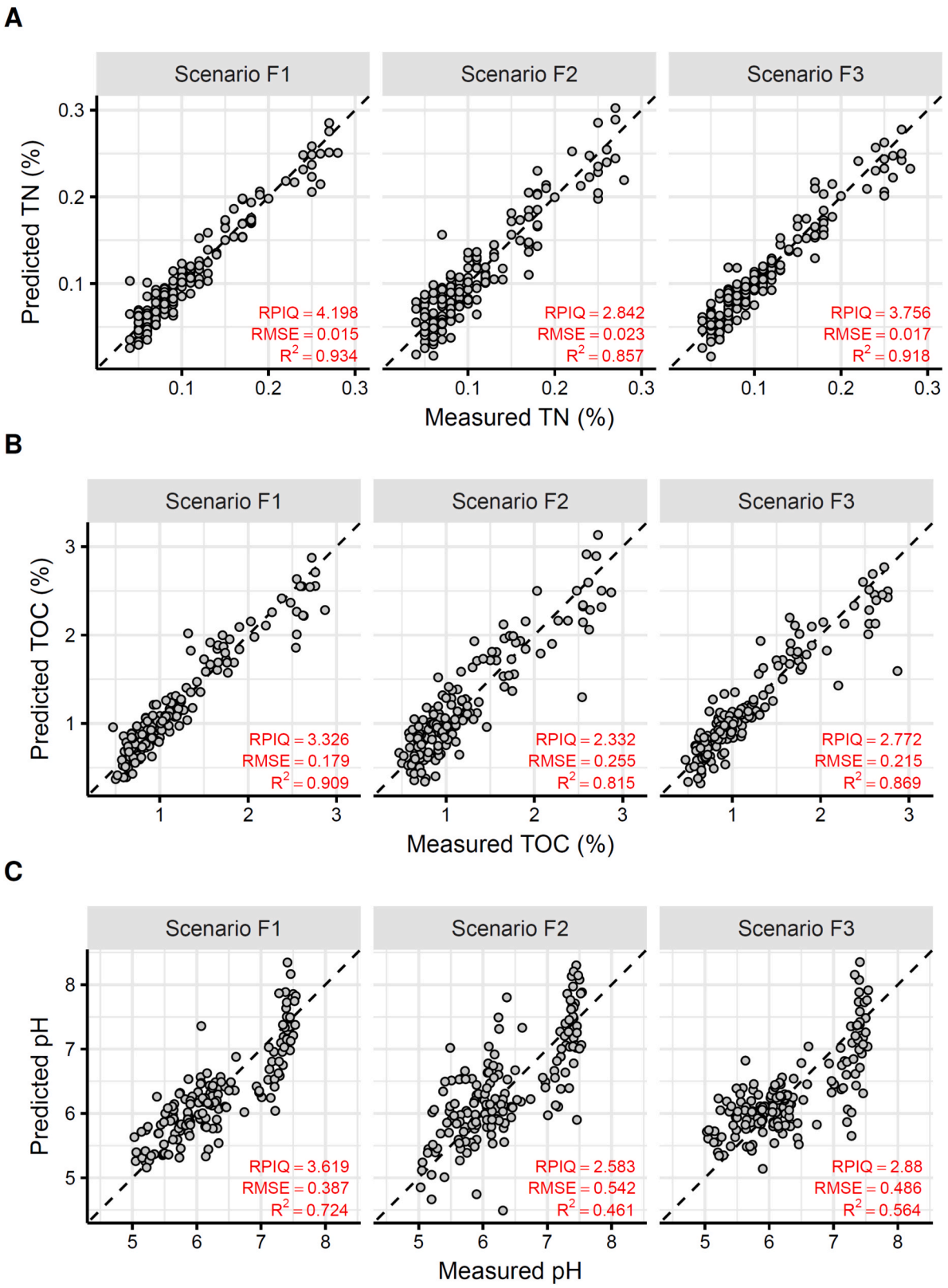


Fig. 9. Prediction performance of the models on the outer-test set under the three field scenarios for A) soil total nitrogen (TN), B) soil total organic carbon (TOC), and C) soil pH (pH). **F1**: Interpolation on the complete dataset as a whole; **F2**: Independent interpolation for the training and test set; **F3**: Independent interpolation within each fold of cross-validation. RPIQ: ratio of performance to interquartile range; RMSE: root mean squared error; R^2: the coefficient of determination. The black and dashed line is the 1:1 line.

results, the predictions on the 10 outer-test sets of the cross-validation were pooled to calculate performance metrics and create scatterplots.

Across all models and field scenarios, the best predicted soil property was TN followed by TOC and pH, respectively. Scenario F1 showed the best performance on the test set, with mean R^2 of 0.93, 0.90, and 0.71 for TN, TOC, and pH, respectively. This was followed by scenario F3 with mean R^2 of 0.91, 0.87, and 0.55 and scenario F2 with mean R^2 of 0.83, 0.78, and 0.46 for TN, TOC, and pH, respectively. In general, the RMSE values of scenario F2 on the test set were about 10 to 30 % higher than those of scenarios F1 and F3. These results illustrate that the performance of the models was different depending on the data leakage scenarios. Comparing the results of lab- and field-based scenarios revealed that they are comparable, especially for TN and TOC, and even in some cases, the field results were even better (Tables 3 and 4). Field-based soil spectroscopy is susceptibility to ambient interferences, and thus, model predictions relying on field spectra are anticipated to exhibit lower performance compared to those based on laboratory spectral analysis (Bönecke et al., 2021; Hong et al., 2018; Kodaira and Shibusawa, 2020; Stenberg et al., 2010; Vogel et al., 2022). This suggests that the relatively good performances of our field scenarios F1 and F2 may be attributed to the impact of data leakage caused by interpolation.

It should also be considered that the only equivalent scenarios between the laboratory and field analyses are scenarios L3 and F3 (Table 2). In both of them, PCA was pipelined and applied to the training set during cross-validation. Furthermore, each data fold under scenario F3 was interpolated independently from other folds. Therefore, the differences in the performance of these scenarios can explain the differences between the predictions of laboratory and field measurements using mobile sensors. As expected, lab scenario L3 had better performances on the test set than field scenario F3. Scenario L3 achieved R^2 (RMSE) values of 0.94 (0.01), 0.88 (0.18), and 0.85 (0.28), while scenario F3 gave the R^2 (RMSE) values of 0.91 (0.02), 0.87 (0.20), and 0.55 (0.48), for TN, TOC, and pH, respectively. The high difference observed for pH is due to the fact that this soil property can be considered much more dynamic compared to TN and TOC (Zhang et al., 2019). Note that the pH values of the soil samples were obtained in December 2020, whereas the sensor measurements were conducted in March and April 2020. Additionally, soil pH does not exhibit a direct spectral response within the Vis-NIR region (Stenberg et al., 2010). Thus, these factors collectively increase the complexity of accurately predicting pH using only a mobile NIR sensor.

In scenario F1, the models achieved the best performances of all field scenarios for all the three soil properties (Table 4 and Fig. 9). This may be mainly due to data leakage as the PCA was pipelined into a PCR model, which helped encapsulating the preprocessing and training steps into the cross-validation. However, due to the interpolation step applied to the entire dataset before splitting, the 10 folds in this scenario may be dependent (as shown in Fig. 5).

Within scenario F2, two independent interpolations were applied on the reflectance values of the training and test sets in each of the 10 outer cross-validation iterations. However, the inner-training and inner-test sets within each inner cross-validation iteration were dependent. They were spatially connected by the interpolation step during the outer iteration (Fig. 6). Thus, the number of components and models were not general enough to predict unseen data of the outer-training set. Furthermore, scenario F2 exhibited the highest discrepancy in model performances between the outer-training and the outer-test sets, with relative differences in R^2 of approximately 14, 20, and 49 % for TN, TOC, and pH, respectively (Table 4). This disparity suggests that the models of scenario F2 may led to stronger overfitting as compared to those of scenarios F1 and F3. Since scenario F2 utilizes independent outer-training and outer-test sets, the high performance observed in scenario F1, where both outer and inner sets are interdependent, can be directly attributed to data leakage. Similarly, in LUCAS scenario 2 (Supplementary File S2), where PCA was applied before data splitting, hyper-parameter tuning was performed on interdependent cross-validation folds, as in scenario F2. This resulted in poor generalization, highlighting the crucial role of truly independent inner-sets for optimal hyper-parameter tuning. Our findings emphasize that even with independent training and test sets, the lack of independent inner-sets for hyper-parameter tuning can lead to suboptimal model performance and poor generalization to unseen data.

In scenario F3, to ensure complete independence between the datasets, separate interpolation steps were applied on each of the 10 cross-validation folds (Fig. 7). Hence, the PCR performance on unseen data increased compared to scenario F2, for all the three soil properties. This improvement, in terms of R^2, was 9, 10, and 16 %, for TN, TOC, and pH, respectively (Table 4 and Fig. 9). This is further illustrated by the non-leaked LUCAS scenario 3 (Supplementary File S2) and the Booßen laboratory scenario L3. By performing model training and hyper-parameter tuning on both independent outer and inner datasets, these non-leaked scenarios yielded models that outperformed their leaked counterparts in predicting the target variables on unseen data. These results highlight the crucial importance of proper pipelining when employing nested cross-validation. This ensures not only the independence of outer-training and outer-test sets but also the independence of inner-training and inner-test sets, effectively mitigating the risk of overfitting.

Upon analyzing Tables 3 and 4, it is apparent that the RMSE values for different soil properties vary depending on the data leakage across the scenarios. The laboratory measurements had higher prediction accuracy as compared to field measurements. In terms of R^2, TN and TOC measurements showed more accurate predictions than pH measurements. Notably, scenario L2 for both TN and TOC measurements had negative R^2 values, which suggests that the models performed worse than the baseline model (using the mean value of the target variable as the prediction). Furthermore, we discovered that our models consistently produced lower RMSE values than the standard deviation of the target soil property, except for scenario L2 (Table 1, Figs. 8 and 9). Hence, it is evident that a model trained under data leakage, such as those in scenario L2, is insufficiently generalizable to make precise predictions on unseen data.

As PCR can be seen as a stacked algorithm in which a PCA is incorporated, it is very easy to encapsulate it with other training workflow steps, such as other preprocessing/feature engineering methods (data scaling, transformation, etc.) and model fitting, into one callable function, also known as pipeline. Thus, a pipeline is a sequence of data preprocessing and modeling steps that can be manipulated as a single unit (Fig. 2). In this way, pipelines help preventing data leakage by ensuring that only those samples from the training set are used for data preprocessing and model fitting (Pedregosa et al., 2011). E.g., when a pipeline with the trained model is applied to the test set, also new and independent preprocessing steps are applied on the test set (Figs. 2-4). To avoid data leakage, we encourage researchers to design leak-free pipelines. These pipelines should encapsulate all preprocessing methods (such as data normalization, scaling, dimensionality reduction, etc.), model training and (if needed) spatial interpolation. A leak-free pipeline ensures that the training and test sets are completely independent of each other.

3.3. Results for spatial autocorrelation (SAC)

Spatial autocorrelation (SAC) may cause dependence between training and test sets used for model calibration, since nearby measurements are dependent (Karasiak et al., 2022).

Randomly assigning samples and measurements to different folds during data splitting can result in nearby samples being placed in different folds, introducing some degree of interdependence between folds.

Models affected by SAC exhibit residuals that are dependent of each other. Therefore, when employing random cross-validation, which does not account for spatial autocorrelation between training and test data, the model's performance might be artificially inflated (Kattenborn et al.,

2022; Ploton et al., 2020).

Therefore, a data splitting design ensuring spatial independence between the training and test sets should be the standard approach for validation of models using spatial data (Karasiak et al., 2022). One possible way to address this issue is to segregate the training and test sets into spatial blocks or clusters during the data-splitting procedure (Meyer et al., 2019; Karasiak et al., 2022). Under this strategy, distance-based buffers around hold-out samples are defined to ensure the learning model uses only spatially independent data. These distances can be determined using a variogram based on the target variable or predictors. This means the minimal spatial distance between the training and test sets should be larger than the largest variogram range obtained among the target variable and predictors. Incorporating spatial dependence (autocorrelation) into the model specifications is another strategy (Hurlbert, 1984; Dormann et al., 2007; Liu et al., 2022).

However, spatial cross-validation can lead to pessimistic map accuracy assessments without notable benefits over standard cross-validation methods, and both approaches have the potential to introduce bias into map accuracy estimates (Wadoux et al., 2021).

Despite using random cross-validation, we generally did not find strong evidence of spatial autocorrelation (SAC) in our results, as assessed through Moran's I on model residuals (see Table 5 and Supplementary Figs. S5 and S6). Notably, the Moran's I values for the actual target soil variables, measured in the soil samples, were significantly higher than those for model residuals. In the plots, this difference appeared as a gradient across the field, with the highest target variable values concentrated in a specific area, while the model residuals exhibited a more random distribution without any clear pattern.

Field scenarios, interestingly, displayed lower Moran's I values and variogram ranges than laboratory scenarios, with F3 exhibiting the lowest Moran's I. This arises from the lower degree of SAC between the outer-training and outer-test sets in F3 compared to other field scenarios, as these sets lacked a common spatial interpolation (Fig. 7). However, it is essential to note that the variogram ranges were typically smaller than the mean distance between sampling points and even smaller than 25 % of the smallest distances. In practical terms, this suggests that in the case of TN and TOC in scenario F3, fewer than four samples per fold might exhibit SAC. Conversely, in scenario F3, pH displayed a notable high absolute Moran's I value, potentially attributed to the model's relatively lower performance, which could amplify the presence of autocorrelation when other sources of data leakage are absent. Consistent with this observation, scenario L2, characterized by PCA applied to independent fold sets without hyper-parameter tuning and displaying the poorest test performance, also exhibited the highest absolute Moran's I values. Additionally, L2 had the largest variogram range, indicating a trend where higher Moran's I values corresponded to larger ranges.

As expected, F1 and F2 (both characterized by spatially dependent folds) exhibited relatively high absolute Moran's I values, highlighting the connection between data leakage, spatial interpolation, and, notably, hyper-parameter tuning (as evidenced in F2), leading to increased spatial autocorrelation (SAC). While we mostly observed no significant SAC in model residuals, we cannot entirely rule it out. Although non-overlapping points among folds might mitigate autocorrelation, inter-fold distances may remain too small to eliminate it.

The lack of significant spatial autocorrelation (SAC) in model residuals could also be attributed to the specific characteristics of our datasets. Supplementary Fig. S1 illustrate the even distribution of measurement points from south to north across a rectangular area with three columns (transects) and 53 rows. This design increases the average distance between sampling points compared to a square or circular layout, helping to reduce spatial dependency between data points when points with large separation distances were selected. By assigning each sampling point to cross-validation folds, the average spatial distance between folds becomes sufficiently large to partially mitigate autocorrelation. While some points in different folds may still exhibit spatial dependence, the majority is spaced far enough apart to minimize this effect.

Further research is needed to comprehensively investigate and manage potential autocorrelation within our dataset. Despite numerous publications addressing this issue, particularly in remote sensing (Kattenborn et al., 2022; Meyer and Pebesma, 2022; Milà et al., 2022; Ploton et al., 2020; Rocha et al., 2018; Wadoux et al., 2021), none, to the best of our knowledge, have addressed it within the context of proximal soil sensing modeling yet.

There is a well-known positive relationship between model complexity and the risk of overfitting (Hawkins, 2004). Highly flexible and overparameterized models, such as polynomial models, may incorporate irrelevant predictors that allow them to fit complex, nonlinear patterns. However, this flexibility can become a liability when

Table 5
Moran's I and variogram parameters for model residuals across laboratory and field scenarios for each soil target variable.

Soil target variable	Scenario*	Moran's I		Variogram**			
		Observed	*p* value	Model	Psill	Range	Kappa
TN	L1	0.0005	0.181	Sph	0.000	46.2	–
	L2	0.0721	0.000	Ste	0.003	74.0	10.0
	L3	−0.0067	0.937	Sph	0.000	48.2	–
	F1	−0.0087	0.649	Ste	0.000	14.0	1.0
	F2	−0.0107	0.396	Sph	0.001	20.1	–
	F3	−0.0032	0.545	Ste	0.000	20.0	0.3
TOC	L1	−0.0046	0.732	Sph	0.010	47.3	–
	L2	0.0609	0.000	Ste	0.310	93.3	1.4
	L3	−0.0052	0.821	Ste	0.005	21.0	10.0
	F1	−0.0102	0.450	Sph	0.033	21.3	–
	F2	−0.0072	0.869	Ste	0.070	9.2	10.0
	F3	−0.0065	0.979	Ste	0.049	11.8	0.7
pH	L1	−0.0005	0.259	Ste	0.064	56.9	1.3
	L2	0.0184	0.000	Gau	0.349	77.9	–
	L3	−0.0048	0.764	Ste	0.043	58.0	0.2
	F1	−0.0079	0.761	Ste	0.027	65.0	10.0
	F2	−0.0054	0.852	Ste	0.302	13.2	0.6
	F3	0.0087	0.003	Ste	0.184	92.9	0.6

* L1: PCA before data splitting and without hyper-parameter tuning; L2: PCA pipelined into a PCR model without hyper-parameter tuning; L3: PCA pipelined into a PCR model with hyper-parameter tuning; F1: Interpolation on the complete dataset as a whole; F2: Independent interpolation for the training and test sets; F3: Independent interpolation within each fold of cross-validation.

** Sph: Spherical model; Ste: Matérn model with M. Stein's Parameterization; Gau: Gaussian model; Psill: partial sill (psill = sill − nugget); Kappa: Smoothness parameter for the Matérn models.

the model begins to capture noise rather than true signal, leading to strong performance on training data but poor generalization to unseen samples. To address complexity-induced overfitting, machine and deep learning algorithms commonly employ regularization techniques such as penalization, dropout, weight decay, pruning, and others (Zou and Hastie, 2005; Srivastava et al., 2014; Manzali and Elfar, 2023; Parhi and Nowak, 2023). In the past year, increasingly complex nonlinear algorithms, such as tree-based methods and deep learning models, have gained popularity in soil sciences (Padarian et al., 2020; Wadoux et al., 2020). However, pipelines prone to overfitting due to high model complexity may be further compromised by data leakage or spatial autocorrelation. It is therefore essential to design pipelines with algorithm-specific regularization parameters carefully. Improper tuning of these settings may result in models that are especially vulnerable to overfitting. Understanding the combined impact of model complexity, data leakage, and spatial autocorrelation effects in soil spectroscopy warrants further investigation.

The concern with leaky pipelines goes beyond overly optimistic performance metrics. By failing to reflect how a model would behave on truly unseen data, they compromise reproducibility, which is essential for the reliable deployment of machine learning models in real-world applications. With growing concerns about reproducibility in the practical use of machine learning-based solutions (Kapoor and Narayanan, 2023), addressing data leakage explicitly contributes not only to methodological rigor but also to the credibility of machine learning applications in agricultural production systems.

4. Conclusions

Despite recognition and study of data leakage in fields such as medicine, genomics, and engineering, no comprehensive survey or study has been specifically addressing this issue in soil science, particularly within soil proximal sensing. This paper aims to illustrate the consequences of data leakage on predicting soil properties using both laboratory- and field-based soil NIR spectroscopy. We explored the risk of data leakage caused by PCA and spatial interpolation during the calibration of models using soil NIR spectra to predict soil properties, including total nitrogen (TN), soil organic carbon (SOC), and pH. To enhance the generalizability of our findings, we further investigated spectral data from soil samples collected across various geographic regions within the European Union, using the LUCAS dataset (Supplementary File S2). This study emphasizes that preprocessing methods, like dimensionality reduction applied to the entire dataset prior to data splitting, compromise the independence of training and test sets. To prevent data leakage, proper pipelining facilitates the independent application of preprocessing methods to training and test sets, thereby ensuring that models generalize effectively to unseen data. The following conclusions are drawn from the findings of this study:

- Applying preprocessing methods such as PCA and spatial interpolation on soil NIR spectral data before data splitting may result in data leakage and model overfitting.
- In a PCR calibration of laboratory-based soil NIR spectra, tuning the optimal number of components (hyper-parameters) is a key to achieve a good performance on the test set. The heuristic determination of the number of components fails to generalize well in unseen data. In particular, hyper-parameter tuning under data leakage generates poor results. Even when training and test sets are independent, hyper-parameter tuning on leaky folds leads to suboptimal parameter values that fail to generalize, ultimately reducing model performance on an independent test set.
- Since field-based spectral data is highly susceptible to interference from external environmental factors and shows a lower signal-to-noise ratio compared to lab-based spectral data, model predictions based on field spectra generally perform worse than laboratory analysis on standardized samples. For *in-situ* soil NIR spectroscopy, data leakage due to spatial interpolation must be considered as an additional factor that can negatively affect the generality of the predictions.
- To avoid data leakage, we encourage researchers to design leak-free pipelines. These pipelines should encapsulate preprocessing methods, model fitting, and if needed spatial interpolation, ensuring that training and test sets are completely independent.
- While significant SAC was generally not observed in model residuals without data leakage, it cannot be entirely ruled out. Despite having non-overlapping points among folds, limited inter-fold distances may still contribute. The model's reduced performance can accentuate SAC when other data leakage sources are absent. Furthermore, under data leakage, hyper-parameter tuning based on dependent folds may exacerbate spatial autocorrelation.

Since data leakage significantly affects model performance across machine learning applications, emphasizing this issue in soil proximal sensing could encourage future studies to refine their methodologies and generate results that are more reliable. For future works, we suggest studies on investigating the risk of data leakage on other types of proximal soil sensing data. It is crucial to be aware of the potential for data leakage when using preprocessing methods on soil spectral data, and to ensure that these methods are only applied after data splitting to prevent overfitting. To evaluate potential data leakage risks, we recommend the following points for consideration:

- Avoid preprocessing before data splitting
- Design leak-free pipelines
  - o Encapsulating preprocessing, model fitting, and spatial interpolation within a pipeline, ensuring it is applied only to training data.
- Ensure proper hyper-parameter tuning
  - o Note that tuning on leaky cross-validation folds produces hyper-parameter values that fail to generalize.
- Avoid interpolation as much as possible
  - o Interpolation is sometimes necessary to achieve spatial alignment of data. Since interpolation relies on information from neighboring data points, it can introduce data leakage and reduce model generalizability.
  - o Interpolation must be done separately for training and test sets.
  - o Consider spatial autocorrelation (SAC) effects. SAC can contribute to data leakage leading to overfitting and artificially inflated performance.
  - o Limited spatial distances between cross-validation folds can contribute to SAC.

CRediT authorship contribution statement

José Correa: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Hamed Tavakoli:** Writing – original draft, Methodology, Investigation, Data curation. **Sebastian Vogel:** Writing – review & editing, Project administration, Investigation, Funding acquisition. **Robin Gebbers:** Writing – review & editing, Supervision, 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.

Acknowledgments

This study was conducted within the BonaRes project "I4S (Intelligence for Soil) – Integrated System for Site-Specific Soil Fertility Management" (https://www.bonares.de/i4s). The authors would like to

acknowledge the Federal Ministry of Education and Research (BMBF) of Germany for funding the I4S project (grant number 031B1069A).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2025.110920.

Data availability

The authors do not have permission to share data.

References

Adamchuk, V.I., Morgan, M.T., Ess, D.R., 1999. An automated sampling system for measuring soil pH. Transactions of the ASAE 42 (4), 885–892.

Alexander, D.L.J., Tropsha, A., Winkler, D.A., 2015. Beware of R^2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J. Chem. Inf. Model. 55 (7), 1316–1322. https://doi.org/10.1021/acs.jcim.5b00206.

Analytical Methods Committee, 2012. Dark uncertainty. Analytical Methods 2012 (4), 2609–2612. https://doi.org/10.1039/C2AY90034C.

Bănică, F.-G., 2012. Chemical Sensors and Biosensors: Fundamentals and applications. John Wiley & Sons, Chichester, UK.

Baumann, P., Helfenstein, A., Gubler, A., Keller, A., Meuli, R.G., Wächter, D., Lee, J., Viscarra Rossel, R., Six, J., 2021. Developing the swiss mid-infrared soil spectral library for local estimation and monitoring. Soil 7, 525–546.

Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J., Elston, D.A., 2010. Regression analysis of spatial data. Ecol. Lett. 13 (2), 246–264.

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML, 2012. International Vocabulary of Metrology – Basic and General Concepts and Associated Terms. 3rd ed. JCGM 200.

Bishop, C., 2006. Pattern recognition and machine learning. Springer, New York.

Bönecke, E., Meyer, S., Vogel, S., Schröter, I., Gebbers, R., Kling, C., Kramer, E., Lück, K., Nagel, A., Philipp, G., Gerlach, F., Palme, S., Scheibe, D., Zieger, K., Rühlmann, J., 2021. Guidelines for precise lime management based on high-resolution soil pH, texture and SOM maps generated from proximal soil sensing data. Precis. Agric. 22 (2), 493–523.

Boumpoulis, V., Michalopoulou, M., Depountis, N., 2023. Comparison between different spatial interpolation methods for the development of sediment distribution maps in coastal areas. Earth Sci. Inf. 16, 2069–2087. https://doi.org/10.1007/s12145-023-01017-4.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132 (3–4), 273–290. https://doi.org/10.1016/j.geoderma.2005.04.025.

Chen, Y., 2013. New Approaches for calculating Moran's Index of Spatial Autocorrelation. PLoS One 8 (7), e68336.

Chen, Y., 2016. Spatial Autocorrelation Approaches to Testing Residuals from Least Squares Regression. PLoS One 11 (1), e0146865.

Corwin, D.L., Lesch, S.M., 2003. Application of soil electrical conductivity to precision agriculture. Agron. J. 95 (3), 455–471.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30 (5), 609–628.

Eckelmann, W., Sponagel, H., Grottenthaler, W., 2005. Bodenkundliche Kartieranleitung.-5. verbesserte und erweiterte-Auflage (Pedological Mapping guidelines. 5th improved and, Extended Edition. Schweizerbart Science Publishers, Stuttgart, Germany.

Ge, Y., Morgan, C.L.S., Wijewardane, N.K., 2020. Visible and near-infrared reflectance spectroscopy analysis of soils. Soil Sci. Soc. Am. J. 84 (5), 1495–1502.

Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta 185, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.

Gebbers, R., 2018. Proximal Soil surveying and monitoring Techniques. In: Stafford, J. (Ed.), Precision Agriculture for Sustainability. Burleigh Dodds Scientific Publishing, Cambridge, UK.

Gia Pham, T., Kappas, M., Van Huynh, C., Nguyen, H.K., L., 2019. Application of ordinary kriging and regression kriging method for soil properties mapping in hilly region of central Vietnam. ISPRS Int. J. Geo Inf. 8 (3), 147.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of Statistical Learning. Data Mining, Inference, and Prediction, Second Edition. Springer, New York.

Hawkins, D.M., 2024. The Problem of Overfitting. Journal of Chemical Information and Computer Sciences 44 (1), 1–12. https://doi.org/10.1021/ci0342472.

Héberger, K., Rácz, A., Bajusz, D., 2017. Which Performance Parameters are best Suited to Assess the Predictive Ability of Models? In: Roy, K. (Ed.), Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences. Springer International Publishing, Cham, pp. 89–104.

Hengl, T., 2009. A Practical Guide to Geostatistical Mapping. University of Amsterdam, Amsterdam.

Hidalgo, D.R., Cortés, B.B., Bravo, E.C., 2021. Dimensionality reduction of hyperspectral images of vegetation and crops based on self-organized maps. Information Processing in Agriculture 8 (2), 310–327.

Hiemstra, P.H., Pebesma, E.J., Twenhöfel, C.J.W., Heuvelink, G.B.M., 2009. Real-time automatic interpolation of ambient gamma dose rates from the dutch radioactivity monitoring network. Comput. Geosci. 35 (8), 1711–1721.

Hong, Y., Yu, L., Chen, Y., Liu, Y., Liu, Y., Liu, Y., Cheng, H., 2018. Prediction of Soil Organic Matter by VIS–NIR Spectroscopy using Normalized Soil Moisture Index as a Proxy of Soil Moisture. Remote Sens. (Basel) 10 (1), 28.

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., Wyble, B., 2020. I tried a bunch of things: the dangers of unexpected overfitting in classification of brain data. Neurosci. Biobehav. Rev. 119, 456–467.

Hurlbert, S.H., 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54 (2), 187–211.

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374 (2065), 20150202.

Joseph, V.R., 2022. Optimal ratio for data splitting. Statistical Analysis and Data Mining: the ASA Data Science Journal 15 (4), 531–538.

Kapoor, S., Narayanan, A., 2023. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4 (9), 100804.

Karasiak, N., Dejoux, J.F., Monteil, C., Sheeren, D., 2022. Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. Mach. Learn. 111 (7), 2715–2740.

Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M.D., Dormann, C.F., 2022. Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. ISPRS Open Journal of Photogrammetry and Remote Sensing 5, 100018.

Kaufman, S., Rosset, S., Perlich, C., 2011. Leakage in data mining: formulation, detection, and avoidance, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. Association for Computing Machinery, San Diego, California, USA, pp. 556–563.

Kodaira, M., Shibusawa, S., 2020. Mobile proximal sensing with visible and near infrared spectroscopy for digital soil mapping. Soil Systems 4 (3), 40.

Legendre, P., 1993. Spatial Autocorrelation: trouble or New Paradigm? Ecology 74 (6), 1659–1673.

Leone, A.P., Viscarra-Rossel, R.A., Amenta, P., Buondonno, A., 2012. Prediction of soil properties with PLSR and vis-NIR spectroscopy: Application to mediterranean soils from Southern Italy. Curr. Anal. Chem. 8 (2), 283–299.

Liu, L., Ji, M., Buchroithner, M., 2017. Combining Partial Least Squares and the Gradient-Boosting Method for Soil Property Retrieval using Visible Near-Infrared Shortwave Infrared Spectra. Remote Sens. (Basel) 9 (12), 1299.

Liu, X., Kounadi, O., Zurita-Milla, R., 2022. Incorporating spatial autocorrelation in machine learning models using spatial lag and eigenvector spatial filtering features. ISPRS Int. J. Geo Inf. 11 (4), 242.

Manzali, Y., Elfar, M., 2023. Random forest pruning techniques: a recent review. Oper. Res. Forum 4, 43. https://doi.org/10.1007/s43069-023-00223-6.

Meyer, H., Pebesma, E., 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nat. Commun. 13 (1), 2208.

Meyer, H, Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecological Modelling 411, 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815.

Milà, C., Mateu, J., Pebesma, E., Meyer, H., 2022. Nearest neighbour distance matching Leave-One-out Cross-Validation for map validation. Methods Ecol. Evol. 13 (6), 1304–1316.

Mitchell, H.B., 2010. Data Fusion: Concepts and ideas, 2 ed. Springer, Berlin, Heidelberg.

Mouazen, A.M., Kuang, B., De Baerdemaeker, J., Ramon, H., 2010. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. Geoderma 158 (1–2), 23–31. https://doi.org/10.1016/j.geoderma.2010.03.001.

Muralidhar, N., Muthiah, S., Butler, P., Jain, M., Yu, Y., Burne, K., Li, W., Jones, D., Arunachalam, P., McCormick, H.S., Ramakrishnan, N., 2021. Using AntiPatterns to Avoid MLOps Mistakes. arXiv:2107.00079.

Nisbet, R., Miner, G., Yale, K., 2018. Handbook of Statistical Analysis and Data Mining applications, 2 ed. Academic Press.

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.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval Partial Least-Squares Regression (iPLS): a Comparative Chemometric Study with an example from Near-Infrared Spectroscopy. Appl. Spectrosc. 54 (3), 413–419. https://doi.org/10.1366/00037020019495.

Nwaila, G.T., Zhang, S.E., Bourdeau, J.E., Frimmel, H.E., Ghorbani, Y., 2024. Spatial interpolation using machine learning: from patterns and regularities to block models. Nat. Resour. Res. 33, 129–161. https://doi.org/10.1007/s11053-023-10280-7.

Padarian, J., Minasny, B., McBratney, A.B., 2020. Machine learning and soil sciences: a review aided by machine learning tools. Soil 6, 35–52. https://doi.org/10.5194/soil-6-35-2020.

Paradis, E., Schliep, K., 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35 (3), 526–528.

Parhi, R., Nowak, R.D., 2023. Deep learning meets sparse regularization: a signal processing perspective. IEEE Signal Process Mag. 40 (6), 63–74. https://arxiv.org/abs/2301.09554.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30 (7), 683–691.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,

Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 11 (1), 4540.
R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online: https://www.R-project.org (accessed on 23 June 2022).
Rinnan, Å., Berg, F.v.d., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry 28(10), 1201-1222.
Rocha, A.D., Groen, T.A., Skidmore, A.K., Darvishzadeh, R., Willemen, L., 2018. Machine Learning using Hyperspectral Data Inaccurately Predicts Plant Traits under Spatial Dependency. Remote Sens. (Basel) 10 (8), 1263.
Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., Scheinost, D., 2024. Data leakage inflates prediction performance in connectome-based machine learning models. Nat. Commun. 15 (1), 1829.
Samala, R.K., Chan, H.-P., Hadjiiski, L., Helvie, M.A., 2021. Risks of feature leakage and sample size dependencies in deep feature extraction for breast mass classification. Med. Phys. 48 (6), 2827–2837.
Schloeder, C.A., Zimmerman, N.E., Jacobs, M.J., 2001. Comparison of methods for interpolating soil properties using limited data. Soil Science Society of America Journal 65 (2), 470–479. https://doi.org/10.2136/sssaj2001.652470x.
Schmidinger, J., Schröter, I., Bönecke, E., Gebbers, R., Ruehlmann, J., Kramer, E., Mulder, V.L., Heuvelink, G.B.M., Vogel, S., 2024a. Effect of training sample size, sampling design and prediction model on soil mapping with proximal sensing data for precision liming. Precis. Agric.
Schmidinger, J., Barkov, V., Tavakoli, H., Correa, J., Ostermann, M., Atzmueller, M., Gebbers, V., S.,, 2024b. Which and how many soil sensors are ideal to predict key soil properties: a case study with seven sensors. Geoderma 450, 117017. https://doi.org/10.1016/j.geoderma.2024.117017.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (56), 1929–1958. https://jmlr.org/papers/v15/srivastava14a.html.
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. Academic Press, pp. 163–215.
Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of soil organic carbon at the european scale by visible and near infrared reflectance spectroscopy. PLoS One 8 (6), e66409.
Tavakoli, H., Correa, J., Sabetizade, M., Vogel, S., 2023. Predicting key soil properties from Vis-NIR spectra by applying dual-wavelength indices transformations and stacking machine learning approaches. Soil Tillage Res. 229, 105684.
Tavakoli, H., Correa, J., Vogel, S., Gebbers, R., 2022. RapidMapper – a mobile multi-sensor platform for the assessment of soil fertility in precision agriculture, VDI Wissensforum(eds.): Proceedings International Conference on Agricultural Engineering (AgEng-LAND.TECHNIK 2022), Berlin, Germany.
Tavakoli, H., Correa, J., Vogel, S., Oertel, M., Zimne, M., Heisig, M., Harder, A., Wruck, R., Pätzold, S., Leenen, M., Gebbers, R., 2024. The RapidMapper: State-of-the-art in mobile proximal soil sensing based on a novel multi-sensor platform. Comput. Electron. Agric. 226, 109443.
Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Econ. Geogr 46, 234–240.
Tziachris, P., Aschonitis, V., Chatzistathis, T., Papadopoulou, M., Doukas, I.J.D., 2020. Comparing machine learning models and hybrid geostatistical methods using environmental and soil covariates for soil pH prediction. ISPRS Int. J. Geo Inf. 9 (4), 276. https://doi.org/10.3390/ijgi9040276.
Viscarra Rossel, R.A., Bouma, J., 2016. Soil sensing: a new paradigm for agriculture. Agr. Syst. 148, 71–74.
Vogel, S., Bönecke, E., Kling, C., Kramer, E., Lück, K., Philipp, G., Rühlmann, J., Schröter, I., Gebbers, R., 2022. Direct prediction of site-specific lime requirement of arable fields using the base neutralizing capacity and a multi-sensor platform for on-the-go soil mapping. Precis. Agric. 23 (1), 127–149.
Wadoux, A.-M.-J.-C., Budiman, M., McBratney, A.B., 2020. Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359. https://doi.org/10.1016/j.earscirev.2020.103359.
Wadoux, A.-M.-J.-C., Heuvelink, G.B.M., de Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. Ecol. Model. 457, 109692.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Yutani, H., 2019. Welcome to the Tidyverse. Journal of Open Source Software 4, 1686.
Wold, S., Johansson, E., Cocchi, M., 1993. PLS: Partial Least Squares Projections to Latent Structures. In 3D QSAR in drug design, Vol. 1, pp. 523-550.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intel. Lab. Syst. 58 (2). https://doi.org/10.1016/S0169-7439(01)00155-1.
Xu, X., Chen, S., Xu, Z., Yu, Y., Zhang, S., Dai, R., 2020. Exploring appropriate preprocessing techniques for hyperspectral soil organic matter content estimation in black soil area. Remote Sens. (Basel) 12 (22), 3765.
Yang, C., Brower-Sinning, R.A., Lewis, G., KÄStner, C., 2023. Data leakage in notebooks: Static detection and better processes, Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering. Association for Computing Machinery, Rochester, MI, USA, Article No.: 30, pp. 1–12.
Ying, X., 2019. An overview of overfitting and its solutions. J. Phys. Conf. Ser. 1168 (2), 022022.
Yli-Halla, M., Schick, J., Kratz, S., Schnug, E., 2016. Determination of Plant Available P in Soil. In: Schnug, E., De Kok, L.J. (Eds.), Phosphorus in Agriculture: 100 % Zero. Springer, Netherlands, Dordrecht, pp. 63–93.
Zhang, Y., Hartemink, A.E., 2020. Data fusion of vis–NIR and PXRF spectra to predict soil physical and chemical properties. Eur. J. Soil Sci. 71 (3), 316–333.
Zhang, Y.-Y., Wu, W., Liu, H., 2019. Factors affecting variations of soil pH in different horizons in hilly regions. PLoS One 14 (6), e0218563.
Zheng, A., Casari, A., 2018. Feature Engineering for Machine Learning. O'Reilly Media Inc, Sebastopol, California, USA.
Zhong, L., Guo, X., Xu, Z., Ding, M., 2021. Soil properties: their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks. Geoderma 402, 115366.
Zhu, J.-J., Yang, M., Ren, Z.J., 2023. Machine learning in environmental research: Common pitfalls and best practices. Environ. Sci. Technol. 57 (46), 17671–17689.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67 (2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.