Geoderma Regional 44 (2026) e01063

Contents lists available at ScienceDirect

Geoderma Regional

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

Rapid analysis of farm-scale soil cadmium concentrations using a regional soil spectral library

G. Shrestha [a,b,*], R. Calvelo-Pereira [a,c], P. Roudier [d,e], G. Kereszturi [a], P. Jeyakumar [a,f], A.P. Martin [g,h], R.E. Turnbull [g], C.W.N. Anderson [a]

[a] Environmental Sciences Group, School of Agriculture and Environment, Massey University, Manawatu Campus, Private Bag 11222, Palmerston North 4442, New Zealand
[b] Bioeconomy Science Institute-AgResearch, Private Bag 11008, Palmerston North 4442, New Zealand
[c] Department of Soils and Natural Resources, Faculty of Agronomy, Universidad de Concepción, Chillán 3812120, Chile
[d] Bioeconomy Science Institute-Manaaki Whenua – Landcare Research, Private Bag 11052, Palmerston North 4442, New Zealand
[e] Te Pūnaha Matatini, A New Zealand Centre of Research Excellence, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[f] Joint FAO/IAEA Division of Nuclear Techniques in Food & Agriculture, International Atomic Energy Agency Head Quarters, Vienna, Austria
[g] Earth Sciences New Zealand, Private Bag 1930, Dunedin 9016, New Zealand
[h] Geological Survey of Victoria, Melbourne 3002, Australia

ARTICLE INFO

Keywords:
Proximal sensing techniques
Potentially toxic trace element
Memory-based learning
Localisation
Environmental monitoring

ABSTRACT

Monitoring soil cadmium (Cd) at farm-scales (average 3 km^2) can potentially be rapid and cost-efficient by implementing proximal sensing techniques benefiting from a leveraged regional-scale (≥ 40,000 km^2) soil spectral library (RSSL). However, prediction models based on RSSL are often of limited use when applied at farm-scales because the coarseness of the RSSL. In this study, a New Zealand RSSL was used to assess the Cd concentration in a farm-scale sample set. For all samples, total Cd was determined, and visible-near-infrared (vis-NIR), mid-infrared (MIR), and portable X-ray fluorescence (pXRF) spectra were collected. A localisation technique to predict farm-scale Cd using RSSL spectral data was developed, based on spectral similarity or land use similarity relative to the farm-scale samples, and/or supplemented with selected farm-scale samples, as input for partial least squares regression and LOCAL algorithms. A model using MIR data from a RSSL pastoral samples subset (n = 283) spiked with 12 extra weighted (×4) farm-scale samples as an input for a LOCAL algorithm, quantified Cd optimally (root mean square error = 0.22 mg Cd/kg; concordance correlation coefficient = 0.78; ratio of performance to interquartile distance = 1.93). Spiking the RSSL subset with farm-scale samples, including otherwise under-represented attributes such as soil order and Cd concentration range, improved the performance of models predicting farm-scale total Cd concentrations. A hybrid technique of localisation approach considered in this study may reduce compliance costs for Cd surveying and management, benefiting farmers.

1. Introduction

Cadmium (Cd) is a trace element in soil that, at elevated plant-available concentrations, can transfer through the food chain potentially causing detrimental effects on ecosystem services, economy, and human wellbeing (Kabata-Pendias, 2010). Soil Cd concentration can increase following volcanic eruptions and human activities including mining and refining activities (Morgan, 2010; Nriagu and Pacyna, 1988). Long-term application of phosphate fertiliser for plant production over decades has accumulated Cd in agricultural soils, amplifying associated risks from farm to fork (Godt et al., 2006; Gray and Cavanagh, 2022; McDowell and Gray, 2022).

Managing soil Cd requires understanding the distribution of Cd at contrasted spatial scales because farmers are interested in determining the ongoing status of Cd in individual farms and paddocks over time (Stahlmann-Brown, 2023), while policy-related assessments of risks

* Corresponding author at: Bioeconomy Science Institute-AgResearch, Private Bag 11008, Palmerston North 4442, New Zealand.
E-mail addresses: gautam.shrestha@agresearch.co.nz (G. Shrestha), rcalvelo@udec.cl (R. Calvelo-Pereira), roudierp@landcareresearch.co.nz (P. Roudier), g.kereszturi@massey.ac.nz (G. Kereszturi), p.jeyakumar@massey.ac.nz (P. Jeyakumar), archie.martin@deeca.vic.gov.au (A.P. Martin), rose.turnbull@dmirs.wa.gov.au (R.E. Turnbull), c.w.n.anderson@massey.ac.nz (C.W.N. Anderson).

https://doi.org/10.1016/j.geodrs.2026.e01063
Received 10 September 2025; Received in revised form 18 February 2026; Accepted 19 February 2026
Available online 20 February 2026

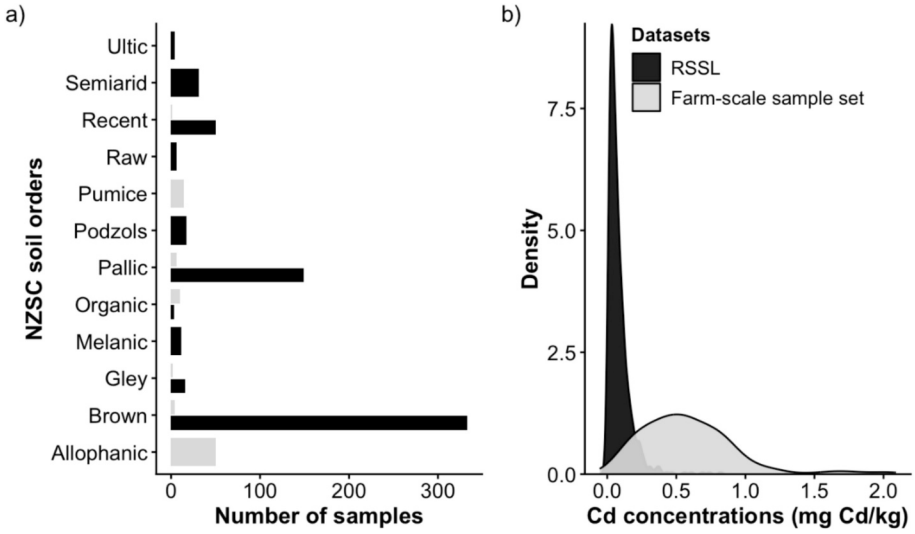

Fig. 1. The distribution of regional-scale soil spectral library (RSSL; black) and farm-scale sample set (grey) shown in a) A histogram of the number of samples per New Zealand soil classification (NZSC) soil orders (following Hewitt (2010)) and b) a density plot of distribution of cadmium (Cd) concentration.

from Cd accumulation require regional-to-national-scale surveys (Ballabio et al., 2024; Bravo et al., 2021). Cadmium monitoring programmes aiming for long-term, repetitive assessment (Tóth et al., 2016) can be expensive in terms of time and resources required for soil sampling campaigns and analysis of Cd in conventional laboratories (Nocita et al., 2015). Deployment of proximal sensors could overcome such limitations by enabling fast, cost-effective, and non-destructive measurement of soil attributes (Nawar et al., 2019; Nduwamungu et al., 2009).

Visible-near-infrared (vis-NIR) and mid-infrared (MIR) spectroscopy effectively assess multiple soil attributes at multiple scales (Di Iorio et al., 2022; Hong et al., 2024; Ma et al., 2024; Nawar and Mouazen, 2017). Within the visible to infrared part of the electromagnetic spectrum, the estimation of Cd relies on its co-variation with spectrally active soil components such as soil organic matter; aluminium (Al) and/or iron (Fe) containing minerals (Soriano-Disla et al., 2014). In contrast, portable X-ray fluorescence (pXRF) based quantification of Cd and other elements in soil is directly proportional to the specific spectral waveband response (Padilla et al., 2019), when the element concentration is above the lower detection limit of the instrument (Rouillon and Taylor, 2016). Cadmium has been quantified using a combination of data from vis-NIR, MIR, and pXRF for regional-scale studies, including from New Zealand (Li et al., 2021; O'Rourke et al., 2016; Shrestha et al., 2022). For farm-scale studies, reliable Cd prediction models have been developed based on vis-NIR only (Kooistra et al., 2001; Shrestha et al., 2024; Zhang et al., 2019). The choice of proximal sensing technique(s) used to assess soil attributes depends on cost, portability, and the precision required (Nawar et al., 2019).

A single soil spectral library (SSL) or multiple SSL, compiling soil spectra and corresponding physical, chemical, biological, and spatial attributes, have been increasingly used in soil science (Viscarra Rossel et al., 2016). There are examples of SSL developed for small, local domains (more similar sites, variation at farm-scale size, e.g., area 1–2000 km^2) and for more diverse and complex domains (comprising the variation over extensive, regional to global, areas, e.g., $\geq$ 30,000 km^2) (O'Rourke et al., 2016; Seidel et al., 2019; Viscarra Rossel et al., 2016). Extensive SSL can be used to build general calibration models for farm-scale applications, reducing costly soil analysis (Brown, 2007; Sila et al., 2016). However, regional- and/or national-scale models are often biased and lack precision when applied at farm-scale, because of, for example, variability in spectral response, land use, geography, climatic conditions, management practices, landscapes, and soil attributes (Gogé et al., 2014; Kuang and Mouazen, 2013; Viscarra Rossel et al., 2024). To overcome such limitations, calibration models based on regional- to national-scale SSL require updating and fine-tuning with information from farm-scale samples. Research to develop such localisation techniques to tailor the models to the specific characteristics of individual sites and produce accurate farm-scale estimates of soil attributes is ongoing (Viscarra Rossel et al., 2024).

Common localisation techniques are similarity-based deterministic methods and spiking (Viscarra Rossel et al., 2024). The deterministic search methods select an optimal subset of samples from the regional- and/or national-scale SSL with similarities to the farm-scale samples, using criteria such as spectral similarity, stratified sampling, and land use type (Lobsey et al., 2017; Shen et al., 2022). Farm-scale prediction using SSL subsets based on one or a few similarity criteria may be of limited use when the SSL includes heterogenous landscapes and diverse soil types, compromising the correlation between spectra and soil attributes (Nocita et al., 2015; Viscarra Rossel et al., 2019). Spiking with farm-scale samples, on the other hand, adds a reasonable number of representative farm-scale samples to the SSL, filling gaps in information (Sankey et al., 2008). Characteristics suggested to select and add farm-scale samples include: (a) spectral similarity, (b) proportional representation of samples (e.g., applying the Kennard-Stone algorithm (Arshad et al., 2021)), and/or (c) filling a concentration gap of a soil attribute of concern (Guerrero et al., 2010). Spiking with extra-weighted (i.e., multiple entries of the same sample) farm-scale samples have also been utilised to enhance representativeness (Greenberg et al., 2022). The regional-scale SSL or its subsets spiked with few farm-scale samples can be economical to improve estimation accuracy only if a small number (5–15%) of farm-scale samples are used (Nocita et al., 2015; Seidel et al., 2019; Viscarra Rossel et al., 2016). Other localisation methods include (i) data-driven heuristic searches, (ii) reusing feature representations, and (iii) combinations of the above (hybrid methods) (Viscarra Rossel et al., 2024). Hybrid methods can improve the accuracy of farm-scale soil attribute estimation using a regional- and/or national-scale SSL by combining the selection of SSL subsets similar to the farm-scale sample set and then addition of key data from farm-scale dataset to complement the information available prior to modelling using algorithms (Viscarra Rossel et al., 2024; Wetterlind and Stenberg, 2010).

Chemometric models relate sample spectra and attribute, implying that the specific algorithm employed for spectral localisation has a major role in generating an optimal output (Brown, 2007; Gogé et al., 2014; Ramirez-Lopez et al., 2013). Partial least squares (PLS) regression is a robust and commonly used when the prediction matrix includes noisy and collinear variables (Li et al., 2020; Nocita et al., 2014). LOCAL, a

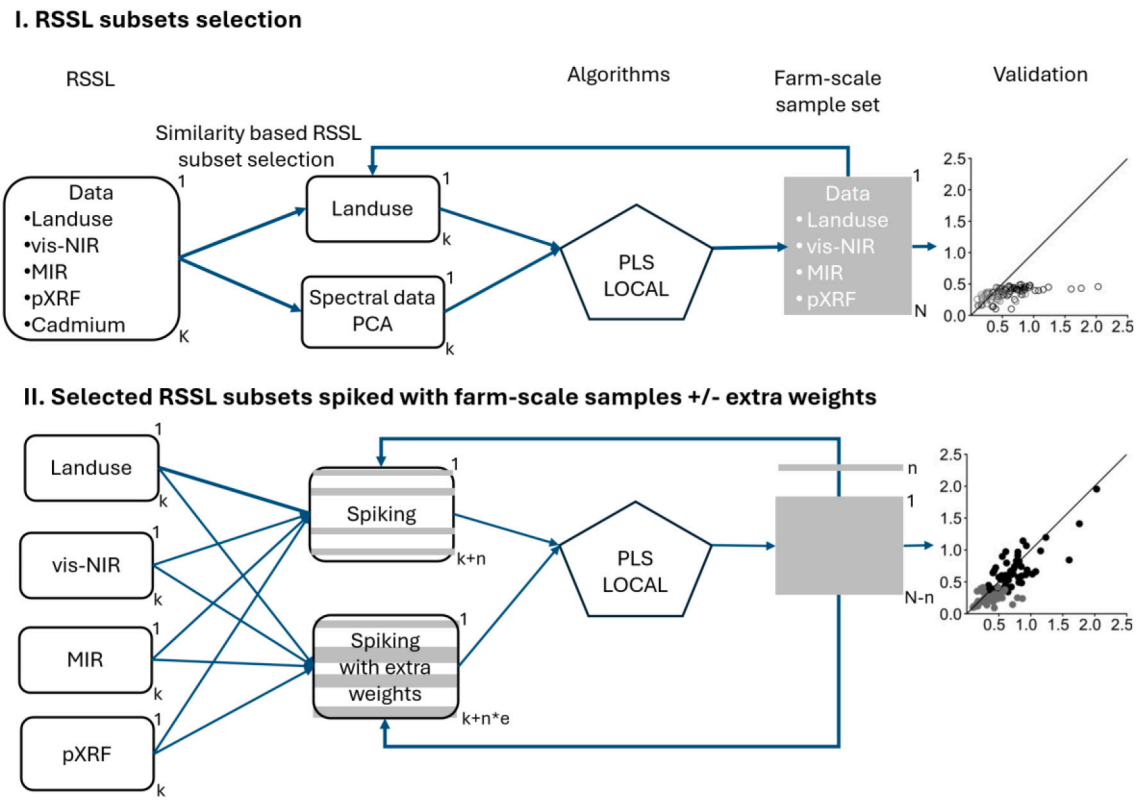

Fig. 2. A workflow diagram showing hybrid spectral localisation technique followed in the study. The regional-scale soil spectral library (RSSL) contains total K (= 625) number of samples and farm-scale set contains total N (= 87) samples. I) RSSL subsets of sample size k were selected based on similarity (principal component analysis) of spectral data (k = 200, 250, 300, 350, 400, 450, 500, 550) of each sensor including visible-near-infrared (vis-NIR), mid-infrared (MIR), or portable X-ray fluorescence (pXRF) or similarity of land use (k = 283) with farm-scale sample set. Partial least square (PLS) regression and LOCAL algorithms were implemented to develop calibration models using individual spectral data to quantify soil cadmium concentrations in farm-scale samples. II) Optimal performing RSSL subsets from land use and each sensor data were spiked with small number of farm-scale samples (n = 6, 12, 18) with or without (+/−) extra weights (e = 4, 7, 9) to optimise the calibration model. The structure of this diagram follows Viscarra Rossel et al. (2024).

memory-based learning algorithm, is also used and develops optimal prediction models specific to each sample using limited computational power (Sanderman et al., 2021; Shenk et al., 1997; Summerauer et al., 2021).

Cadmium monitoring in agricultural farm soils, essential to manage the risks associated with this potentially toxic trace element, could be improved using legacy regional-scale SSL (Ng et al., 2022; Shrestha et al., 2022). This study utilises a legacy, regional-scale SSL containing vis-NIR, MIR, and pXRF spectroscopic information as well as laboratory-derived Cd concentration results, to predict Cd levels in farm-scale samples. The objective of the study is to develop a comprehensive workflow for farm-scale Cd prediction using regional-scale SSL by implementing combination of spectral localisation techniques involving (1) three individual proximal sensor data, (2) similarity-based deterministic search methods based on attributes as spectral similarity and land use similarity, then (3) spiking with farm-scale samples with or without extra weights, and (4) two algorithms (PLS regression or LOCAL algorithm). The results of this study, developing a hybrid method enabling efficient use of regional-scale dataset for cost-effective estimation of farm-scale Cd with a reasonable accuracy, will be useful to researchers interested in monitoring farm-scale Cd distribution and policymakers involved in regional-scale decision making.

2. Materials and methods

2.1. Regional-scale soil spectral library and farm-scale sample set

Two datasets were used in this study: a regional-scale SSL (RSSL) comprising samples from a baseline soil geochemical survey in the South Island of New Zealand (Martin et al., 2016; Rattenbury et al., 2018; Shrestha et al., 2022) and a farm-scale set containing samples from pastoral soils (Shrestha et al., 2024).

The RSSL represents topsoil (0–20 cm, n = 625) samples collected at a regular 8 km spacing covering c. 40,000 km^2 between sea level and 2000 m above sea level within the Otago and Southland regions of New Zealand (Table S1; Martin et al. (2016); Rattenbury et al. (2018); Shrestha et al. (2022)). The most common (45%) land use in the RSSL was pasture (Table S1). Prevalent soil orders were Brown (53%) and Pallic (24%) soils as per the New Zealand Soil Classification (Hewitt (2010); Fig. 1a; Table S1). The measured soil Cd concentration ranged between 0.005 and 1.31 mg Cd/kg, with an average of 0.08 mg Cd/kg in the RSSL (Fig. 1b; Table S1).

The farm-scale (average farm size: 3 km^2; range: 1–1800 km^2 in NZ MfE (2021); Pāmu (2025)) set includes topsoil (0–15 cm, n = 87) samples collected from representative pastoral (dairy, sheep, and beef) farms in the Waikato, Canterbury, and Southland regions of New Zealand with a history of long-term application of Cd-rich phosphate fertiliser (see Fig. S1 in Shrestha et al. (2024); Stafford (2017)). Allophanic (57%) and Pumice (16%) soils predominated this dataset (Fig. 1a; Table S1). For this study, to visualise the effect of soil types on Cd estimation, samples were divided into two distinct groups: Allophanic that is characterised by Al-rich aluminosilicate clay minerals and high phosphorus retention capacity, and non-Allophanic that includes all other soils (Hewitt, 2010). In the farm-scale samples set, the soil Cd concentration ranged between 0.10 and 2.03 mg Cd/kg, with an average of 0.58 mg Cd/kg (Fig. 1b; Table S1).

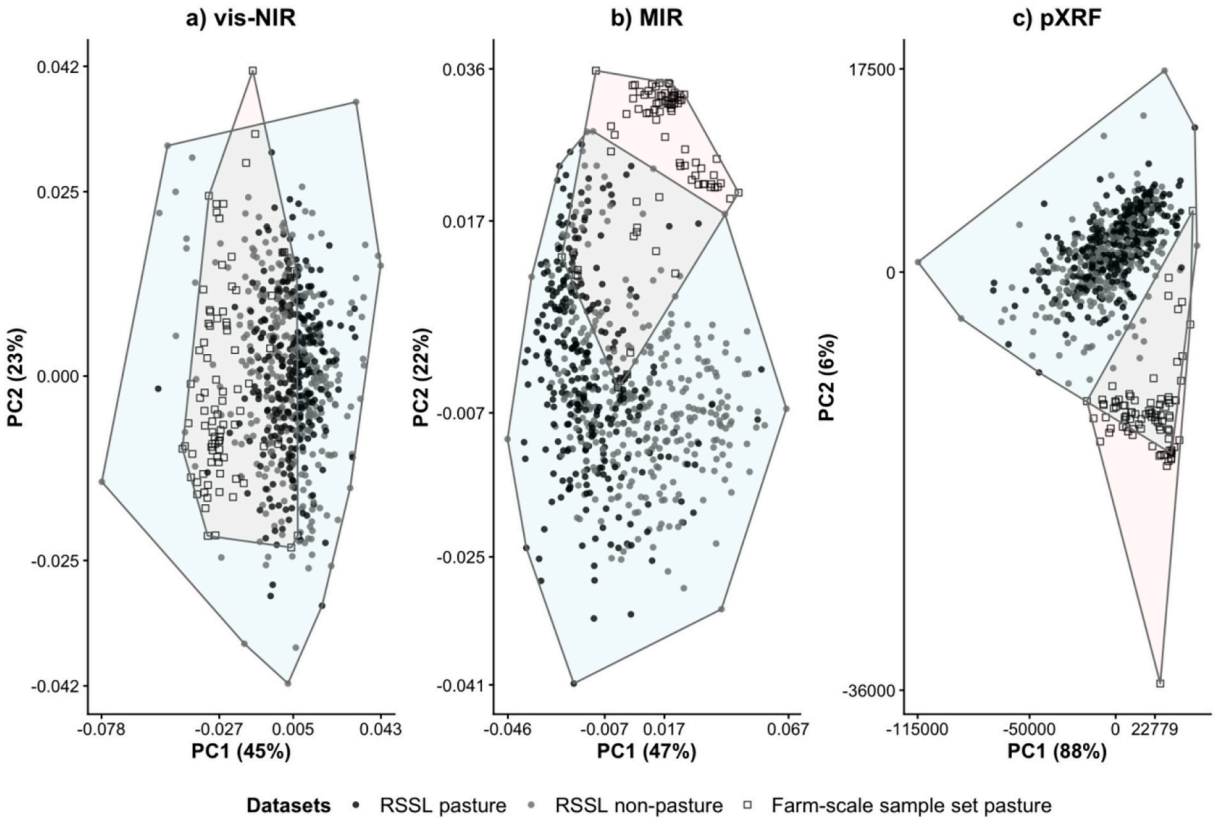

Fig. 3. Principal component analysis (PCA) of (a) visible-near-infrared (vis-NIR), (b) mid-infrared (MIR), and (c) portable X-ray fluorescence (pXRF) spectra. Data are shown for the regional-scale soil spectral library (RSSL; $n = 625$), including pasture (black dots) and non-pasture (grey dots) samples, and the farm-scale sample set ($n = 87$; transparent rectangles). Spectral variance explained by two main principal components (PC1 and PC2) are shown as percentage (%) value within the brackets in both axes. The RSSL hull is shaded in pale green polygon and farm-scale sample set hull is shaded in pale pink polygon.

2.1.1. Sample preparation, spectral data collection, and pre-processing

For all soils, air-dried representative subsamples were scanned using three proximal sensors: vis-NIR, MIR, and pXRF. Visible-NIR (350–2500 nm) reflectance spectra were recorded using an ASD FieldSpec3 spectroradiometer (Analytical Spectral Devices Inc., Boulder, Colorado, USA) fitted with a contact probe containing a 4.5 W halogen bulb as a light source. For scanning, the probe was pushed in contact with the sub − 2 mm samples filled in Petri dishes to a 10 mm depth (Shrestha et al., 2024; Shrestha et al., 2022).

Mid-infrared (7498–600 cm^{-1}) diffuse reflectance spectra were captured using a Fourier transformed infra-red (FTIR) spectrometer (Vertex 70, Bruker, Germany) equipped with a microplate reader extension for high throughput screening infrared spectroscopy equipment (HTS-XT, Tensor II, Bruker, Germany). For scanning, an aluminium microtitre plate with 48 wells and vertical gutters separating each line of 8 wells were packed with sub − 0.5 mm soil (4 wells per sample) prepared by grinding each sample in a Retsch RM200 mortar grinder for 30 s.

Pre-processing of vis-NIR spectra included splice correction, removal of the noisiest part of the spectra (350–399 nm) and then pseudo-absorbance ($\log_{10}(R)$) transformation. Both vis-NIR and MIR spectra were derived and smoothed using a Savitzky-Golay filter (first-order derivative, window size = 9, and second-order polynomial) (Savitzky and Golay, 1964).

The pXRF spectra (0–40 keV) were measured using an Olympus Vanta C series instrument with a rhodium anode and used in *Geochem* mode (*2 beams*). A cylindrical plastic cup (22 mm height by 14 mm diameter) was packed with sub − 2 mm soil and covered with 4 mm thick polypropylene film prior to each measurement (Shrestha et al., 2022). Raw spectra (*beam 1*) were used after removing the first 24 wavelength data with zero spectral response values.

The distribution of multi-dimensional spectral data was assessed following principal component analysis (PCA). Spectra from each sensor was mean centred before performing the PCA. For the combined dataset of RSSL and farm-scale sample set, maximum possible 712 principal components were computed. The distribution of all samples in the PCA space was plotted using the first two principal components describing the greatest amount of variance in the datasets.

2.2. Developing a spectral localisation technique to predict Cd concentration in farm-scale samples

In this study, a hybrid spectral localisation technique to predict Cd was developed in two steps, using vis-NIR, MIR, and pXRF data independently (Fig. 2). The first step was to develop Cd prediction models using the entire RSSL set and/or subsets selected by similarity of either spectra or land use with farm-scale sample set. In the second step, the selected RSSL subsets from the first step were spiked with a few farm-scale samples with or without extra weights (Fig. 2). Subsequently, selected spectra and Cd concentration data were used as input for chemometric modelling (PLS regression, LOCAL algorithm) (Fig. 2). The models developed were evaluated by comparing their performance. All chemometric analyses were performed in the R statistical environment (RStudio Team, 2021). Total Cd concentration values were log-transformed, and then mean centring and variance scaling were performed, prior to use in the predictive modelling.

2.2.1. Step 1: selection of RSSL subsets

The RSSL subsets were selected based on two similarity criteria: (i) spectra and (ii) land use (Fig. 2). Samples from the RSSL spectrally alike to the samples in the farm-scale set were selected using PCA of each proximal sensor data (vis-NIR, MIR, or pXRF). The Mahalanobis distance

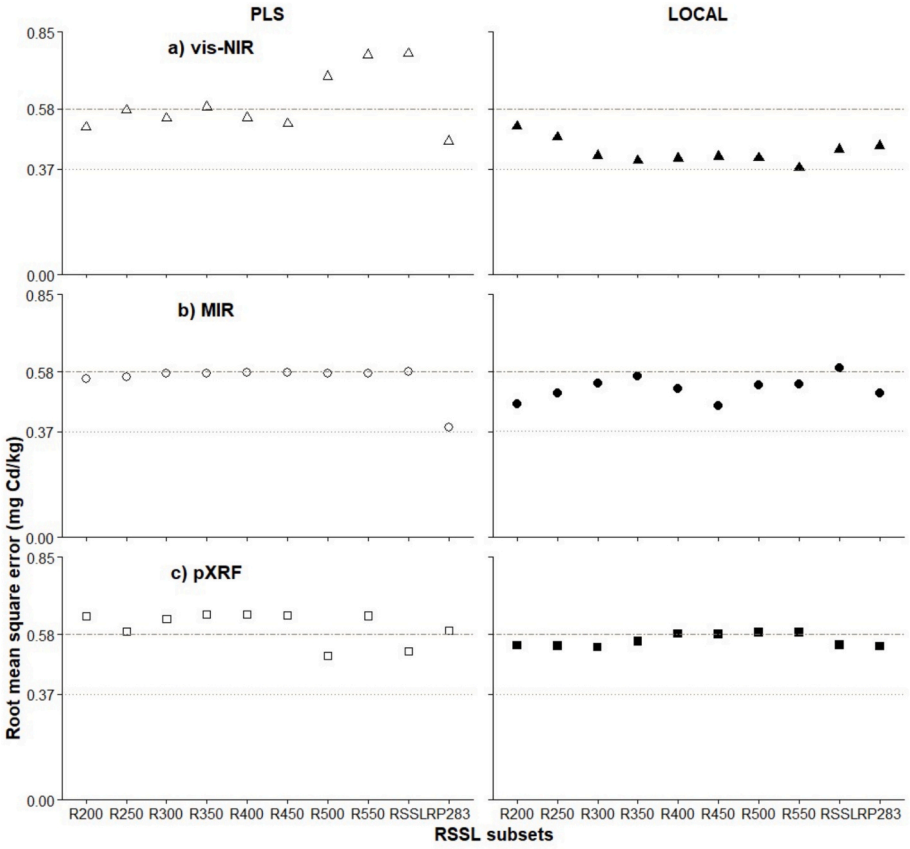

Fig. 4. Dot plots showing the root mean square error of cadmium (Cd) concentrations (mg Cd/kg) prediction for farm-scale samples (n = 87) using regional-scale soil spectral library (RSSL) subsets (R200, R250, R300, R350, R400, R450, R500, or R550), selected by principal component analysis of (a) visible-near-infrared (vis-NIR, triangles), (b) mid-infrared (MIR, circles), and (c) portable X-ray fluorescence (pXRF, rectangles) data, used as input for partial least squares (PLS) regression (hollow shapes) or LOCAL algorithm (black coloured shapes). The RSSL and RSSL pastoral samples subset (RP283) were included for comparison. The dashed line shows the average Cd concentration (0.58 mg Cd/kg) of the farm-scale sample set. The dotted line represents the minimum root mean square error value (0.37 mg Cd/kg) achieved by predictive modelling.

values of a RSSL sample from each farm-scale sample were computed using R package *resemble* (Ramirez Lopez et al., 2016) and summed. The RSSL samples with the least-sum distance value to the farm-scale sample set were considered the most similar. Consequently, the RSSL (R625) was clustered into subsets: R200, R250, R300, R350, R400, R450, R500, and R550 (Fig. 2). A RSSL subset of all samples from pastoral (P) land use (a total of 283 samples) was included (RP283; Fig. 2).

2.2.2. *Step 2: spiking with farm-scale samples with or without extra weights*

Selected RSSL subsets (assessed as optimal from Step 1) were independently spiked with selected farm-scale samples (Fig. 2). The selection of farm-scale samples was based on: (1) a representative set of 20% of the farm-scale samples separated using the Kennard-Stone algorithm (Kennard and Stone, 1969) termed L18, (2) within the subset of 20% samples, 66% were chosen, including soil orders absent in the RSSL i.e., Allophanic and Pumice soils. This selection was termed L12, and (3) within the subset of 20% samples, 33% of samples were chosen that contained Cd concentrations absent in the RSSL. This selection was termed L6. Extra-weights (Greenberg et al., 2022) of four (×4), seven (×7), or nine (×9) were given to the spiked farm-scale samples (L18; L12; L6) (Fig. 2).

2.2.3. *Modelling framework*

The R package *pls* (Mevik et al., 2020) was used for PLS regression wrapped in the *caret* package (Kuhn et al., 2021). During model development, repeated cross-validation (10 folds, twenty-five repeats) was used for PLS hyper-parameterisation. The optimal number of latent variables retained in each model were chosen based on the root mean square error (RMSE), following one standard error of the empirically optimal model (Breiman et al., 1984).

The R package *resemble* (Ramirez Lopez et al., 2016) was used for LOCAL algorithm. For LOCAL algorithm, dissimilarity thresholds were set between 0.01 and 1 at 0.01 increments. The minimum size allowed for the neighbourhood was 80 and the maximum was set at the total number of samples in the training set. This algorithm selects unique samples to use as predictors, validated by the nearest neighbour. Multiple regression cross-validation (i.e., between a minimum of four and a maximum of 25 components) were used to develop regression models, which were then weighted and averaged to obtain the final predicted value (Shenk et al., 1997). All prediction models developed using RSSL subsets and spiked RSSL subsets were evaluated using 10-fold cross-validation on the farm-scale sample set.

2.2.4. *Model performance assessment*

The predictive accuracy of the different models was assessed and compared (see details in Shrestha et al., 2022) calculating RMSE, coefficient of determination (R^2), ratio of performance to interquartile distance (RPIQ; Bellon-Maurel et al. (2010)), Lin's concordance correlation coefficient (CCC; Lin (1989)), and bias. Additionally, PLS loadings, linear combination of predictors maximising covariance with a target property (Kuhn and Johnson, 2013), generated for the optimal models were used to interpret the prediction results. The PLS loadings were plotted for RSSL, farm-scale sample set, selected RSSL subsets, and those RSSL subsets spiked with farm-scale samples with or without extra

Table 1
Validation results of prediction models using proximal sensors: visible-near-infrared (vis-NIR), mid-infrared (MIR), or portable X-ray fluorescence (pXRF) data of I) regional-scale soil spectral library (RSSL) and selected RSSL subsets (optimal performing sets each featuring either land use similarity (RP283) or spectral similarity based on principal component analysis of vis-NIR (R550), MIR (R450), or pXRF (R500) and II) selected RSSL subsets from first step spiked with farm-scale samples (L, n = 6, 12, 18) with or without (+/−) extra weights (×4, ×7, ×9) as input for partial least squares (PLS) regression or LOCAL algorithm predicting total soil cadmium (Cd) concentrations in farm-scale samples.

Proximal sensors	Datasets	PLS					LOCAL				
		RMSE	R^2	CCC	RPIQ	Bias	RMSE	R^2	CCC	RPIQ	Bias
I: RSSL and subset selection based on similarity of spectra and land use											
vis-NIR		0.78	0.31	0.33	0.56	0.43	0.44	0.16	0.37	0.99	−0.16
MIR		0.58	0.00	0.00	0.75	−0.46	0.59	0.00	0.00	0.73	−0.47
pXRF	RSSL	0.52	0.04	0.16	0.83	−0.25	0.54	0.11	0.07	0.80	−0.43
vis-NIR		0.47	0.30	0.47	0.93	0.16	0.45	0.08	0.28	0.96	−0.05
MIR		**0.38**	**0.33**	**0.21**	**1.13**	**−0.24**	0.51	0.05	0.07	0.86	−0.37
pXRF	RP283	0.59	0.11	0.03	0.73	−0.49	0.54	0.02	0.03	0.81	−0.41
	R450	0.53	0.33	0.46	0.82	0.22	0.41	0.28	0.46	1.05	−0.19
vis-NIR	R500	0.69	0.30	0.37	0.62	0.30	0.41	0.28	0.47	1.06	−0.17
	R550	0.77	0.29	0.34	0.56	0.35	**0.37**	**0.32**	**0.53**	**1.16**	**−0.12**
	R450	0.58	0.00	0.00	0.75	−0.46	**0.46**	**0.17**	**0.16**	**0.94**	**−0.33**
MIR	R500	0.57	0.00	0.00	0.75	−0.46	0.53	0.02	0.05	0.81	−0.40
	R550	0.57	0.00	0.00	0.76	−0.45	0.54	0.04	0.07	0.81	−0.40
	R450	0.65	0.11	0.01	0.67	−0.55	0.58	0.10	0.04	0.75	−0.48
pXRF	R500	**0.50**	**0.04**	**0.14**	**0.86**	**−0.31**	0.59	0.06	0.03	0.74	−0.48
	R550	0.64	0.11	0.01	0.67	−0.55	0.59	0.06	0.04	0.74	−0.48
II: Selected RSSL subsets spiked with farm-scale samples +/− extra weights											
vis-NIR		0.39	0.31	0.53	1.12	0.01	0.31	0.34	0.58	1.39	−0.04
MIR		0.27	0.39	0.55	1.60	0.02	0.25	0.55	0.72	1.76	−0.05
pXRF	RP283 + L6	0.46	0.09	0.16	0.95	−0.3	0.38	0.21	0.45	1.15	−0.07
vis-NIR		0.32	0.36	0.60	1.34	−0.01	0.28	0.48	0.66	1.57	−0.10
MIR		0.27	0.39	0.52	1.58	−0.01	0.23	0.64	0.75	1.87	−0.10
pXRF	RP283 + L12	0.22	0.46	0.65	0.87	−0.03	0.33	0.31	0.51	1.31	−0.13
vis-NIR		0.37	0.27	0.51	1.18	−0.01	0.28	0.43	0.63	1.56	−0.07
MIR		0.28	0.38	0.47	1.53	−0.06	0.24	0.57	0.72	1.78	−0.08
pXRF	RP283 + L18	0.44	0.09	0.17	0.98	−0.28	0.34	0.30	0.49	1.28	−0.14
vis-NIR		0.33	0.33	0.57	1.31	−0.04	0.26	0.53	0.68	1.65	−0.10
MIR		0.27	0.42	0.52	1.60	−0.03	**0.22**	**0.66**	**0.78**	**1.93**	**−0.08**
pXRF	RP283 + L12 × 4	0.38	0.12	0.29	1.14	−0.14	0.35	0.25	0.46	1.23	−0.14
vis-NIR		0.31	0.38	0.61	1.38	−0.06	0.28	0.51	0.66	1.57	−0.12
MIR		0.28	0.41	0.54	1.53	−0.09	0.25	0.60	0.74	1.76	−0.10
pXRF	RP283 + L12 × 7	0.38	0.12	0.31	1.14	−0.13	0.34	0.30	0.46	1.26	−0.16
vis-NIR		0.31	0.40	0.62	1.42	−0.07	0.27	0.52	0.67	1.60	−0.11
MIR		0.27	0.44	0.59	1.58	−0.09	0.23	0.63	0.77	1.86	−0.09
pXRF	RP283 + L12 × 9	0.38	0.11	0.31	1.13	−0.11	0.34	0.32	0.48	1.27	−0.17
	R550 + L6	0.54	0.31	0.45	0.81	0.16	0.32	0.44	0.59	1.34	−0.17
vis-NIR	R550 + L12	0.43	0.28	0.49	1.00	0.01	**0.27**	**0.59**	**0.69**	**1.60**	**−0.15**
	R550 + L18	0.41	0.31	0.53	1.05	0.05	0.31	0.44	0.59	1.42	−0.15
	R450 + L6 × 7	0.32	0.20	0.38	1.35	−0.05	0.29	0.45	0.66	1.47	−0.04
MIR	R450 + L12 × 7	0.30	0.35	0.57	1.42	−0.08	0.24	0.59	0.76	1.80	−0.06
	R450 + L18 × 7	0.29	0.43	0.62	1.49	−0.10	**0.24**	**0.62**	**0.77**	**1.83**	**−0.07**
	R500 + L6	0.48	0.09	0.14	0.89	−0.35	0.35	0.24	0.49	1.25	0.00
pXRF	R500 + L12	0.45	0.09	0.21	0.97	−0.26	**0.32**	**0.31**	**0.52**	**1.35**	**−0.11**
	R500 + L18	0.44	0.08	0.21	0.99	−0.23	0.35	0.30	0.49	1.24	−0.16

Optimal model performance parameters based on land use and each sensor data are in bold letters. Performance statistics parameters include root mean square error (RMSE), co-efficient of determination (R^2), concordance correlation co-efficient (CCC), ratio of performance to interquartile distance (RPIQ), and bias.

weights for the visual assessment.

3. Results

3.1. Spectral characteristics of RSSL and farm-scale datasets

Fig. 3 shows the relative distribution of the RSSL and farm-scale datasets in the space defined by the first two principal components calculated from the (a) vis-NIR, (b) MIR, and (c) pXRF data. The first two principal components explain 68, 69, and 94% of the spectra variation for vis-NIR, MIR, and pXRF, respectively. In PC1 and PC2 space, the RSSL and farm-scale samples set show almost complete overlap for the vis-NIR data whereas the spectral similarity is less pronounced for MIR, and pXRF datasets (Fig. 3).

3.2. Predicting Cd in farm-scale samples using RSSL subsets

The models based on the RSSL subsets based on spectral similarity show a limited ability to predict Cd in farm-scale samples (Fig. 4; Tables 1 and S2). The best model for Cd prediction was based on vis-NIR data from 550 samples (R550; Fig. 4) using LOCAL algorithm with a RMSE value of 0.37 mg/kg soil (CCC = 0.53; RPIQ = 1.16; Fig. 4a and Tables 1 and S2).

The prediction models developed from RSSL subsets selected based on land use (i.e., RSSL pastoral samples subset; RP283;) outperformed (relatively consistent RMSE, 0.38–0.59 mg Cd/kg) models developed from spectral similarity (RMSE: 0.37–0.77 mg Cd/kg; Fig. 4; Tables 1 and S2). Among the three proximal sensors, the model using MIR data from the RP283 subset as input for PLS regression quantified Cd in farm-scale sample set with RMSE of 0.38 mg Cd/kg soil (CCC = 0.21, RPIQ = 1.13; Figs. 4 and 6a; Tables 1 and S2).

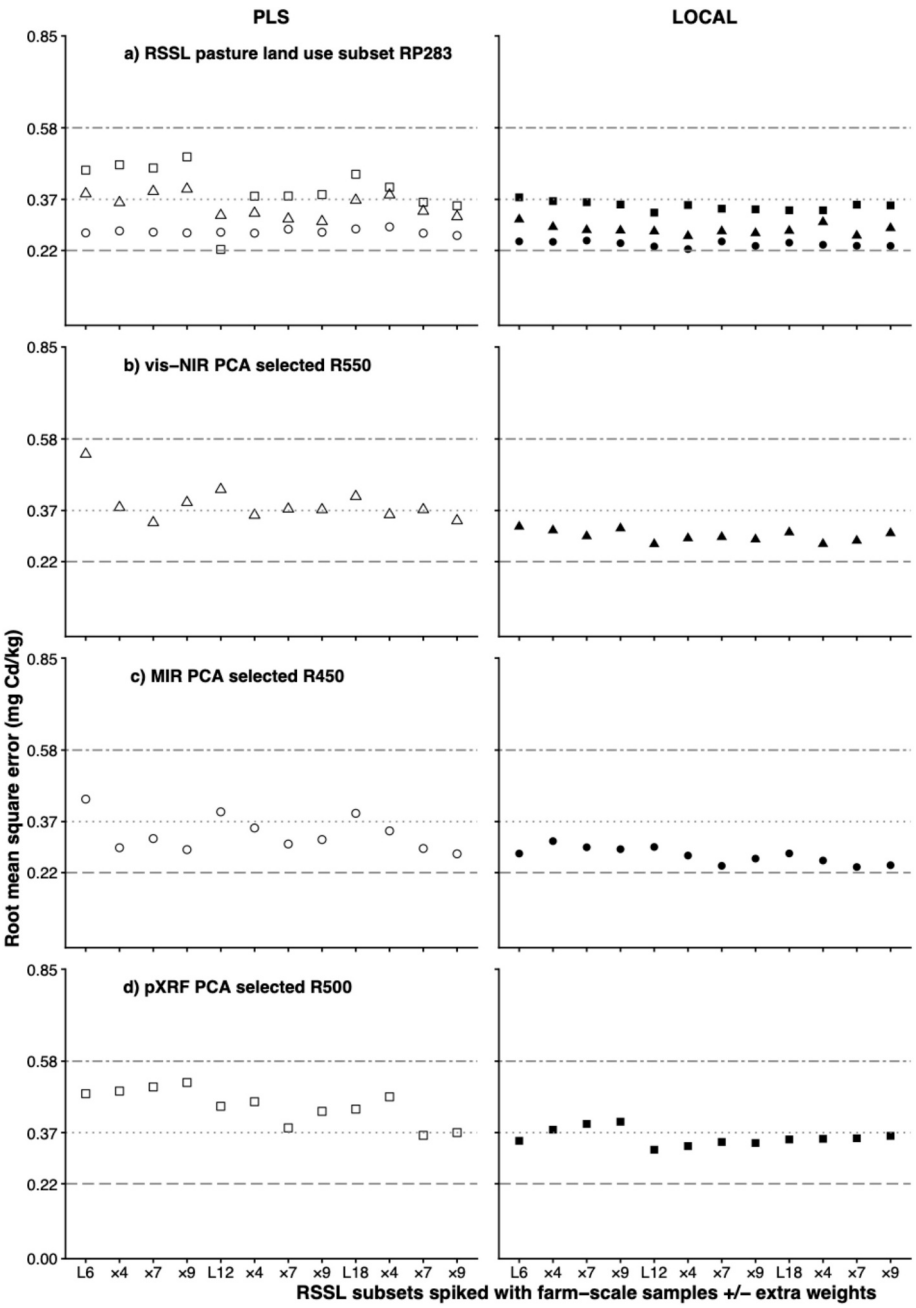

Fig. 5. Dot plots showing the root mean square error of cadmium (Cd) concentrations (mg Cd/kg) predictions for farm-scale samples (n = 87) using regional-scale soil spectral library (RSSL) subsets by a) visible-near-infrared (vis-NIR, triangles), mid-infrared (MIR, circles), or portable X-ray fluorescence (pXRF, rectangles) data and using a) RSSL pastoral samples subset RP283, b) vis-NIR principal component analysis (PCA) selected R550, c) MIR PCA selected R450, or d) pXRF PCA selected R500, spiked with farm-scale samples (L6, L12, or L18) with or without (+/−) extra weights (×4, ×7, or ×9) as input for partial least squares (PLS) regression (hollow shapes) and LOCAL algorithm (black coloured shapes). The dash-dot line is the average Cd concentration (0.58 mg Cd/kg) in the farm-scale sample set. The dotted line is the minimum root mean square error (RMSE = 0.37 mg Cd/kg) achieved by predictive modelling using only RSSL subsets. The dashed line is the minimum RMSE (0.22 mg Cd/kg) obtained by predictive modelling using RSSL subset spiked with farm-scale samples with or without (+/−) extra weights.

3.3. Predicting farm-scale Cd using spiked RSSL subsets

The prediction models based on RSSL subsets spiked with farm-scale samples outperformed those models based solely on RSSL subsets (Figs. 5 and 6). Particularly, models using LOCAL algorithm outperformed those based on PLS regression when spiked RSSL subsets were used (Fig. 5; Tables 1 and S3). LOCAL models were less influenced by the extra weights (≈5% variation in RMSE) given to the included farm-scale samples, whereas PLS models were severely influenced by extra weights (>10% variation in RMSE; Fig. 6; Tables 1 and S3).

Localised models quantifying farm-scale Cd based on MIR data outperformed those models based on vis-NIR data (Figs. 5 and 6; Tables 1 and S3). The model based on MIR data from RSSL pastoral subset RP283, spiked with 12 extra weighted (×4) farm-scale samples (i.e., RP283 + L12 × 4) as input for the LOCAL algorithm, quantified Cd optimally (RMSE = 0.22 mg Cd/kg; CCC = 0.78; RPIQ: 1.93) (Figs. 5 and 6e).

Prediction models using MIR data to predict soil Cd using R450 subset (MIR PCA) or pastoral subset (RP283) spiked with farm-scale samples as input for the LOCAL algorithm performed with low RMSE (0.22–0.31 mg Cd/kg soil; CCC = 0.63–0.78; Fig. 5; Table S3).

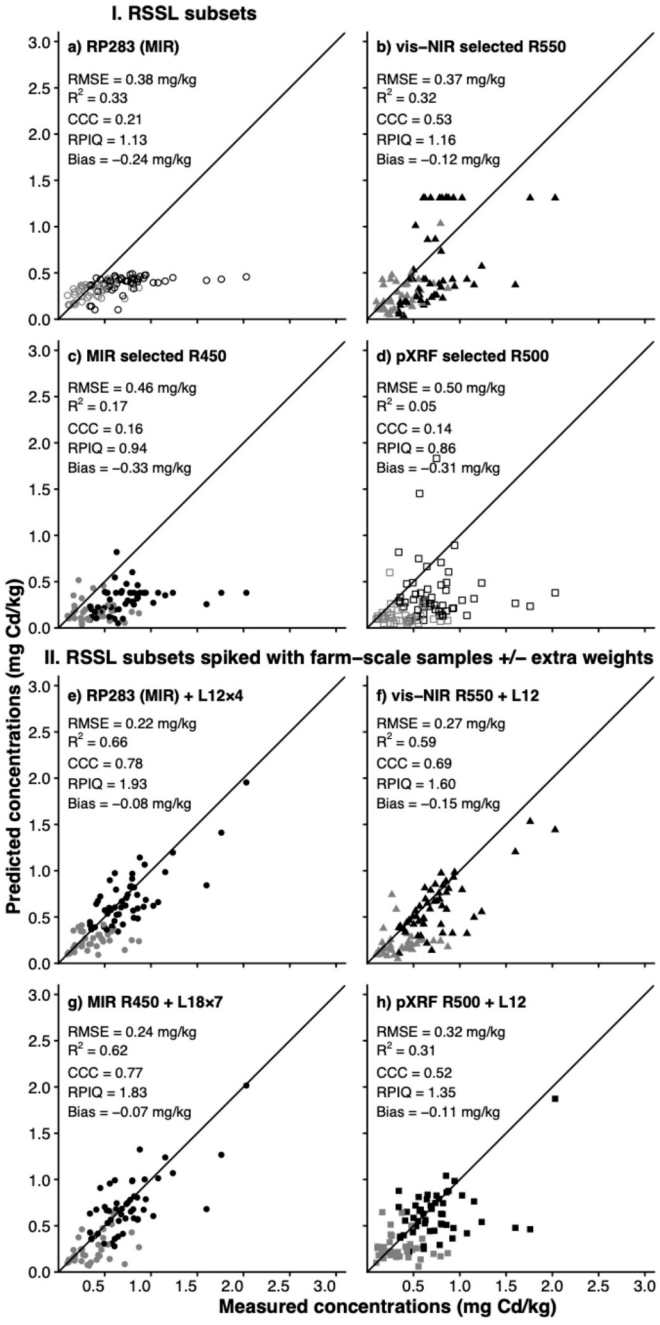

Fig. 6. Measured versus predicted cadmium (Cd) concentrations (mg Cd/kg) for the farm-scale sample set (n = 87) based on optimal calibration models using I) selected RSSL subsets a) RSSL pastoral samples subset RP283 (MIR spectra) and RSSL subsets of spectral similarity (selected based on principal component analysis) b) vis-NIR, c) MIR, and d) pXRF and II) optimal performing RSSL subsets, selected from I, spiked with farm-scale samples with or without (+/−) extra-weights for e) pastoral subset (using MIR data, circles) and each proximal sensor data: f) visible-near-infrared (vis-NIR, triangle), g) mid-infrared (MIR, circles), and h) portable X-ray fluorescence (pXRF, rectangles) as input for partial least squares (PLS) regression (hollow shapes) or LOCAL algorithm (black coloured shapes). Soil samples are coloured black for Allophanic and grey for non-Allophanic soils. Performance statistics parameters: root mean square error (RMSE), co-efficient of determination (R^2), concordance correlation co-efficient (CCC), ratio of performance to interquartile distance (RPIQ), and bias.

Prediction models using vis-NIR data to predict soil Cd after spiking R550 subset (vis-NIR PCA) or pastoral subset (RP283) as input for the LOCAL algorithm performed with RMSE of 0.26–0.32 mg Cd/kg soil (CCC = 0.58–0.69; Fig. 5; Table S3). In general, models based on pXRF data performed poorly (RMSE = 0.32–0.40 mg Cd/kg soil; CCC = 0.35–0.52 Fig. 5d; Tables S2 and S3).

3.4. *Important wavelengths*

The PLS loadings highlight the relative importance of certain spectral regions which helped to explain the performance of a particular prediction model (Fig. 7). For Cd estimation, loadings showed the importance of (1) 1300–1450 nm (Al- and Fe-clay minerals), 1800–2000 nm (clay minerals, soil organic matter), and 2200–2500 nm (soil organic matter–mineral complex) regions of the vis-NIR spectra; and (2) 2130–1700 cm^{-1} (metal–carbonyl (-CO) groups in soil organic matter) and 3700–3000 cm^{-1} (minerals, soil organic matter) regions of the MIR spectra (Fig. 7).

4. Discussion

4.1. *Farm-scale soil Cd concentration prediction*

This study shows that a hybrid spectral localisation technique leveraging the useful but coarse information from a RSSL (samples collected at a regular 8 km spacing, covering an area c. 40,000 km^2) is potentially applicable to develop models quantifying total soil Cd concentration at farm-scales. The comprehensive workflow (Fig. 2) relies on the robust combination of similarity-based deterministic search methods, leading to proximal sensor data selection, optimal subset selection, and then spiking with farm-scale samples with or without extra weights (Fig. 2). This workflow leveraged RSSL information to assess farm-scale Cd concentrations (Fig. 6). These findings are encouraging for fast and indirect estimation of soil Cd using vis-NIR and MIR proximal sensors, despite the inherent complexity in predicting soil Cd and other trace elements (Kooistra et al., 2001; Wu et al., 2010). The Cd associated with spectrally active soil components (e.g., functional groups in soil organic matter; reactive Al and Fe; Fig. 7) is indirectly quantified because both soil organic matter and minerals, of variable spatial distribution, control Cd speciation, transfer, and accumulation over time (Janik et al., 1998; Niazi et al., 2015; Wang et al., 2017). Thus, RSSL can potentially be repurposed to assess farm-scale samples, actively reducing analysis costs (Nawar et al., 2019; Shepherd and Walsh, 2002). The results highlight the relevance of developing SSL including regional/national scale pedological diversity (Shen et al., 2022; Viscarra Rossel et al., 2022), particularly when aligned with complementary studies such as soil geochemical baselines (Martin et al., 2016).

The hybrid technique developed to localise spectroscopic modelling involving multiple sensor data, similarity based search methods, spiking with farm-scale samples with or without extra weights, and algorithms that have been used independently, usually one at a time (Mendes et al., 2022; Moura-Bueno et al., 2020). The potential to accurately predict Cd concentration depends on the coverage and diversity of the RSSL, in combination with the similarity between the regional- and farm-scale datasets (Figs. 1 and 3; Viscarra Rossel et al. (2022)). Here, key features were (i) land use or spectral similarity (as evaluated by PCA; Fig. 3; Gogé et al. (2014)) of the RSSL subset and (ii) spiking with extra weighted small (14%) proportion of farm-scale samples representing Allophanic and Pumice soil orders and with Cd concentrations of > 0.60 mg Cd/kg soil; selected using the Kennard-Stone algorithm. Overall, the localisation technique efficiently addressed key information gaps (pedological diversity, Cd concentration) in the calibration model (Brown, 2007; Guerrero et al., 2010; Wetterlind and Stenberg, 2010).

The localised prediction models based on LOCAL algorithm were less influenced by extra weights given to the spiked farm-scale samples (Figs. 5 and 6; Tables 1, S2 and S3). The LOCAL algorithm selects similar

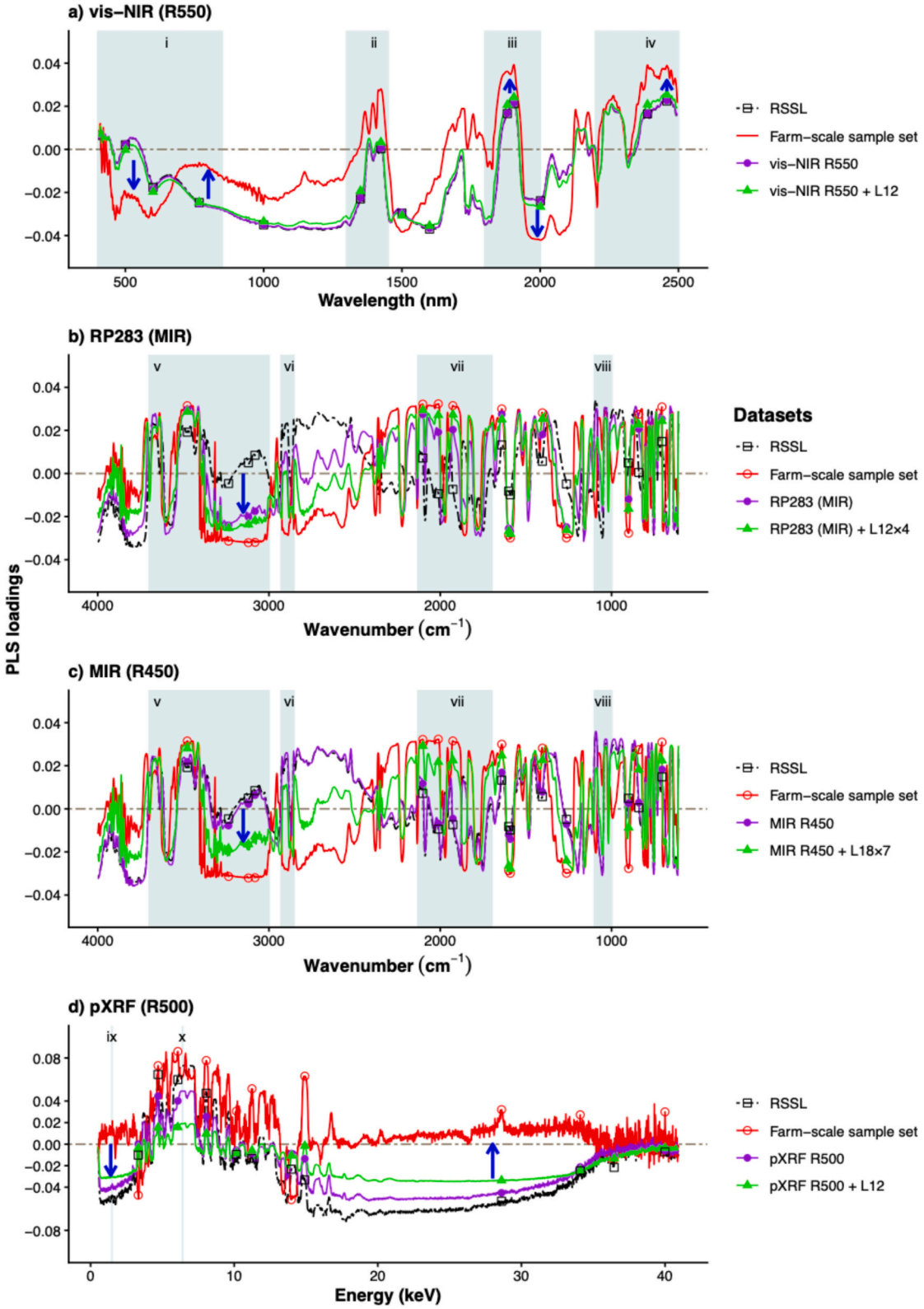

Fig. 7. Partial least squares (PLS) loadings for the regional-scale soil spectral library (RSSL, black), farm-scale sample set (red), and selected RSSL subsets (pink) a) R550, b) RP283, c) R450, or d) R500 and their best performing spiked with farm-scale samples with or without extra-weights (green). The shaded areas: i) soil colour, iron oxides, and soil organic matter (400–850 nm), ii) minerals, water related, and – OH bond (1300–1450 nm), iii) water bonded with Al– and Fe– containing minerals and soil organic matter (1800–2000 nm), iv) Hydroxyl bonded with Al– and Fe– containing minerals and soil organic matter (2200–2500 nm), v) Fe– and Al– containing minerals (3700–3000 cm^{-1}), vi) alkyl (2929–2855 cm^{-1}), vii) metal–carbonyl (2130–1700 cm^{-1}), viii) quartz (1100–1000 cm^{-1}), ix) Al (1.48 keV), and x) Fe (6.40 keV). Arrows show the regions where PLS loadings for the RSSL and farm-scale sample set differ.

but unique sample spectra and discards duplicates (Shenk et al., 1997), which may constitute an advantage when developing localised models (Dangal et al., 2019).

In this study, selecting representative samples with common land use in both datasets improved the accuracy of localised Cd prediction models to some extent (i.e., RP283; Fig. 6a). This outcome is comparable to Moura-Bueno et al. (2020) who found land use-based selection of a vis-NIR SSL subset for soil organic carbon estimation in farm-scale samples is more accurate than other selection methods, such as geographic regions, soil texture class, or spectral similarity. Selecting an optimal number of samples of similar land use ensured the balanced inclusion of spectrally similar soils and comparable Cd concentrations in this study (Figs. 1, 3, and 7b).

4.2. Modelling constraints

Models developed using solely RSSL subsets, by themselves, were of limited application (Figs. 1, 4, and 6ad; Tables 1 and S1). Differences in Cd concentration range, soil orders, soil depth, and geography between the two datasets constrained accuracy of models based on RSSL or its subsets for farm-scale sample Cd prediction.

First, the overall low Cd concentrations in the RSSL set (average = 0.08 mg Cd/kg soil; Fig. 1b) than in the farm-scale sample set (average = 0.58 mg Cd/kg soil; Fig. 1b), resulted inaccurate predictions, notably for pXRF based models, despite spiking with extra weighted farm-scale samples (Lemière, 2018; Weindorf and Chakraborty, 2020). Second, differences in carbon concentration, which influences Cd distribution – among other attributes (Kooistra et al., 2001; Minasny et al., 2006), and changes notably with depth, impact model prediction due to a dilution effect. In this sense, RSSL samples, collected from 0 to 20 cm depth, showed relatively low total carbon (4.1%) than the farm-scale samples, collected from 0 to 15 cm depth (total carbon: 9.5%) (Kooistra et al., 2001; Minasny et al., 2006; Shrestha et al., 2024; Shrestha et al., 2022). Third, soil order may also influenced model predictions, Brown soils (53% of samples in the RSSL) typically show low carbon concentration (total carbon: 4.3%) in contrast to Allophanic soils (58% of samples in the farm-scale sample set; total carbon: 8.5%) (Table S1; Shrestha et al. (2024); Shrestha et al. (2022)).

These differences between the two datasets may have conditioned the predictive performance of models based on vis-NIR data to quantify Cd (Figs. 6b and f; Tables 1 and S3). Overall, this is due to the increase in noise from multiple overtones associated with the fundamental vibrations of molecular bonds and functional groups (Siebielec et al., 2004; Stenberg and Rossel, 2010).

In this study, the geographical overlap between two datasets was low (RSSL included only the 20% of farm-scale samples), and this dissimilarity, affecting the distribution of multidimensional spectral data (Fig. 3), could constrain similarity-based search and spiking and thus reduce model performance (Moura-Bueno et al., 2020). Geographical mismatch was indirectly associated to differences in soil orders in this study, as indicated above.

4.3. Future work

The main findings in this study suggest that building a comprehensive regional−/national- scale SSL could further contribute to a cost-effective assessment of multiple soil attributes at the farm-scale using proximal sensing techniques (Ng et al., 2022). There are country- and continent- scale SSL available in public domains (e.g., Mendes et al. (2022); Viscarra Rossel et al. (2016)), with information collected using proximal sensors with different specifications, requiring calibration transfer before they can be widely applied to estimate soil attributes (Pittaki-Chrysodonta et al., 2021; Sanderman et al., 2021). Investigating how a systematic assessment of geographical superposition between regional- and/or national-scale and farm-scale datasets, especially considering archived and legacy SSL (Nocita et al., 2015; Viscarra Rossel et al., 2016) would help to enhance model generalisation capacity and improve comparability among datasets deserves attention. Further development of (pedological-) transfer functions correcting variability (e.g., soil moisture and texture, matrix effect, pedological-diversity) is also needed (Minasny et al., 2011).

Significant and continuous investment to develop and/or use extensive SSL representing soil variations at multiple scales will improve environmental characterisation at local scales. It will further create the possibility of using of regional-scale SSL to monitor farm-scale soil variations and contaminants in near real-time, using, for example airborne and spaceborne hyperspectral remote sensing (Kuang and Mouazen, 2013; Wang et al., 2022). Further, with the integration of machine learning with one or more proximal sensor data, similarity based deterministic search methods, and algorithms can make real-time predictions of soil attributes.

5. Conclusions

Regional-scale SSL have successfully been used here to quantify soil Cd in farm-scale pastoral soils by developing a hybrid spectral localisation technique, involving a comprehensive workflow for datasets including three proximal sensors data, spectral and land use similarity based deterministic search methods for subset selection, spiking with farm-scale samples with or without extra weights, and comparing two algorithms. This study, relative to earlier studies, shows the need to combine different similarity based, deterministic search methods to optimise calibration models. Developing national-scale SSL covering major pedological variations, land uses, and soil attributes of concern can make farm-scale sample characterisation rapid and inexpensive. Leveraging extensive SSL may help improve land management practices relevant to farmers, researchers, and policy makers.

CRediT authorship contribution statement

G. Shrestha: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis. **R. Calvelo-Pereira:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization. **P. Roudier:** Writing – review & editing, Validation, Supervision, Methodology, Data curation, Conceptualization. **G. Kereszturi:** Writing – review & editing, Supervision. **P. Jeyakumar:** Writing – review & editing, Supervision. **A.P. Martin:** Writing – review & editing, Resources. **R.E. Turnbull:** Writing – review & editing, Resources. **C.W.N. Anderson:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgements

Gautam Shrestha was supported by Massey University School of Agriculture and Environment scholarship, Helen E Akers postgraduate scholarship, Pūtea Tuatoko – doctoral financial support grant, and Kathleen Spragg agricultural research trust award for his doctoral study. Authors acknowledge Mark Rattenbury, and other personnel at Earth Sciences New Zealand (formerly GNS Science), who contributed to the collection and analysis of soil survey samples from the Otago and Southland regions. We are grateful for the comments of Brendan Malone (CSIRO, Australia) and Reddy Pullanagari (The University of Adelaide, Australia) on a draft of this paper. We thank two anonymous reviewers for their helpful comments to improve the manuscript and Prof. Ingrid Kögel-Knabner for editorial handling.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geodrs.2026.e01063.

Data availability

Data will be made available on request.

References

Arshad, M., Zhao, D., Zare, E., Sefton, M., Triantafilis, J., 2021. Proximally sensed digital data library to predict topsoil clay across multiple sugarcane fields of Australia: applicability of local and universal support vector machine. Catena 196, 104934. https://doi.org/10.1016/j.catena.2020.104934.

Ballabio, C., Jones, A., Panagos, P., 2024. Cadmium in topsoils of the European Union – an analysis based on LUCAS topsoil database. Sci. Total Environ. 912, 168710. https://doi.org/10.1016/j.scitotenv.2023.168710.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. Trends Anal. Chem. 29, 1073–1081. https://doi.org/10.1016/j.trac.2010.05.006.

Bravo, D., Leon-Moreno, C., Martínez, C.A., Varón-Ramírez, V.M., Araujo-Carrillo, G.A., Vargas, R., Quiroga-Mateus, R., Zamora, A., Rodríguez, E.A.G., 2021. The first National Survey of cadmium in cacao farm soil in Colombia. Agronomy 11, 761. https://doi.org/10.3390/agronomy11040761.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Routledge.

Brown, D.J., 2007. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma 140, 444–453. https://doi.org/10.1016/j.geoderma.2007.04.021.

Dangal, S.R.S., Sanderman, J., Wills, S., Ramirez-Lopez, L., 2019. Accurate and precise prediction of soil properties from a large mid-infrared spectral library. Soil Syst. 3, 11. https://doi.org/10.3390/soilsystems3010011.

Di Iorio, E., Napoletano, P., Circelli, L., Memoli, V., Santorufo, L., De Marco, A., Colombo, C., 2022. Comparison of natural and technogenic soils developed on volcanic ash by Vis-NIR spectroscopy. Catena 216, 106369. https://doi.org/10.1016/j.catena.2022.106369.

Godt, J., Scheidig, F., Grosse-Siestrup, C., Esche, V., Brandenburg, P., Reich, A., Groneberg, D.A., 2006. The toxicity of cadmium and resulting hazards for human health. J. Occup. Med. Toxicol. 1, 22. https://doi.org/10.1186/1745-6673-1-22.

Gogé, F., Gomez, C., Jolivet, C., Joffre, R., 2014. Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? Geoderma 213, 1–9. https://doi.org/10.1016/j.geoderma.2013.07.016.

Gray, C.W., Cavanagh, J.A.E., 2022. The state of knowledge of cadmium in New Zealand agricultural systems: 2021. N. Z. J. Agric. Res. 66 (4), 285–335. https://doi.org/10.1080/00288233.2022.2069130.

Greenberg, I., Seidel, M., Vohland, M., Koch, H.-J., Ludwig, B., 2022. Performance of in situ vs laboratory mid-infrared soil spectroscopy using local and regional calibration strategies. Geoderma 409, 115614. https://doi.org/10.1016/j.geoderma.2021.115614.

Guerrero, C., Zornoza, R., Gomez, I., Mataix-Beneyto, J., 2010. Spiking of NIR regional models using samples from target sites: effect of model size on prediction accuracy. Geoderma 158, 66–77. https://doi.org/10.1016/j.geoderma.2009.12.021.

Hewitt, A.E., 2010. New Zealand Soil Classification: Manaaki-Whenua - Landcare Research, New Zealand.

Hong, Y., Sanderman, J., Hengl, T., Chen, S., Wang, N., Xue, J., Zhuo, Z., Peng, J., Li, S., Chen, Y., Liu, Y., Mouazen, A.M., Shi, Z., 2024. Potential of globally distributed topsoil mid-infrared spectral library for organic carbon estimation. Catena 235, 107628. https://doi.org/10.1016/j.catena.2023.107628.

Janik, L.J., Merry, R.H., Skjemstad, J.O., 1998. Can mid infrared diffuse reflectance analysis replace soil extractions? Aust. J. Exp. Agric. 38, 681–696. https://doi.org/10.1071/ea97144.

Kabata-Pendias, A., 2010. Trace Elements in Soils and Plants. CRC Press. https://doi.org/10.1201/b10158.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137–148. https://doi.org/10.2307/1266770.

Kooistra, L., Wehrens, R., Buydens, L.M.C., Leuven, R.S.E.W., Nienhuis, P.H., 2001. Possibilities of soil spectroscopy for the classification of contaminated areas in river floodplains. Int. J. Appl. Earth Obs. Geoinf. 3, 337–344. https://doi.org/10.1016/s0303-2434(01)85041-8.

Kuang, B., Mouazen, A.M., 2013. Effect of spiking strategy and ratio on calibration of on-line visible and near infrared soil sensor for measurement in European farms. Soil Tillage Res. 128, 125–136. https://doi.org/10.1016/j.still.2012.11.006.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3.

Kuhn, M., Wing, J., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., R Core Team, Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., 2021. Caret: Classification and Regression Training. R package version 6.0–90.

Lemière, B., 2018. A review of pXRF (field portable X-ray fluorescence) applications for applied geochemistry. J. Geochem. Explor. 188, 350–363. https://doi.org/10.1016/j.gexplo.2018.02.006.

Li, H., Jia, S., Le, Z., 2020. Prediction of soil organic carbon in a new target area by near-infrared spectroscopy: comparison of the effects of spiking in different scale soil spectral libraries. Sensors (Basel, Switzerland) 20, 4357. https://doi.org/10.3390/s20164357.

Li, F., Xu, L., You, T., Lu, A., 2021. Measurement of potentially toxic elements in the soil through NIR, MIR, and XRF spectral data fusion. Comput. Electron. Agric. 187, 106257. https://doi.org/10.1016/j.compag.2021.106257.

Lin, L.I.K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268. https://doi.org/10.2307/2532051.

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

Ma, Y., Minasny, B., Roudier, P., Theng, B.K.G., Carrick, S., 2024. Application of mid-infrared (MIR) spectroscopy to identify and quantify minerals in New Zealand soils. Catena 242, 108115. https://doi.org/10.1016/j.catena.2024.108115.

Martin, A.P., Turnbull, R.E., Rattenbury, M.S., Cohen, D.R., Hoogewerff, J., Rogers, K.M., Baisden, W.T., Christie, A.B., 2016. The regional geochemical baseline soil survey of southern New Zealand: design and initial interpretation. J. Geochem. Explor. 167, 70–82. https://doi.org/10.1016/j.gexplo.2016.05.009.

McDowell, R.W., Gray, C.W., 2022. Do soil cadmium concentrations decline after phosphate fertiliser application is stopped: a comparison of long-term pasture trials in New Zealand? Sci. Total Environ. 804, 150047. https://doi.org/10.1016/j.scitotenv.2021.150047.

Mendes, W.D.S., Demattê, J.A.M., Rosin, N.A., Terra, F.D.S., Poppiel, R.R., Urbina-Salazar, D.F., Boechat, C.L., Silva, E.B., Curi, N., Silva, S.H.G., José Dos Santos, U., Souza, Valladares G., 2022. The Brazilian soil mid-infrared spectral library: the power of the fundamental range. Geoderma 415, 115776. https://doi.org/10.1016/j.geoderma.2022.115776.

Mevik, B.-H., Wehrens, R., Hovde, K., Liland, P.H., 2020. pls: partial least squares and principal component regression. R package version 2.7–3.

MfE, 2021. Farm Numbers and Size, 2002–2019. Ministry for the Environment via MfE Data Service. https://data.mfe.govt.nz/table/105409-farm-numbers-and-size-2002-2019/.

Minasny, B., McBratney, A.B., Mendonça-Santos, M.L., Odeh, I.O.A., Guyon, B., 2006. Prediction and digital mapping of soil carbon storage in the lower Namoi Valley. Aus. J. Soil Res. 44, 233–244. https://doi.org/10.1071/sr05136.

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

Morgan, R., 2010. Soil, heavy metals, and human health. In: Brevik, E.C., Burgess, L. (Eds.), Soils and Human Health. CRC Press, pp. 59–82.

Moura-Bueno, J.M., Dalmolin, R.S.D., Horst-Heinen, T.Z., Ten Caten, A., Vasques, G.M., Dotto, A.C., Grunwald, S., 2020. When does stratification of a subtropical soil spectral library improve predictions of soil organic carbon content? Sci. Total Environ. 737, 139895. https://doi.org/10.1016/j.scitotenv.2020.139895.

Nawar, S., Mouazen, A.M., 2017. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. Catena 151, 118–129. https://doi.org/10.1016/j.catena.2016.12.014.

Nawar, S., Cipullo, S., Douglas, R.K., Coulon, F., Mouazen, A.M., 2019. The applicability of spectroscopy methods for estimating potentially toxic elements in soils: state-of-the-art and future trends. Appl. Spectrosc. Rev. 55, 525–557. https://doi.org/10.1080/05704928.2019.1608110.

Nduwamungu, C., Ziadi, N., Parent, L.-É., Tremblay, G.F., Thuriès, L., 2009. Opportunities for, and limitations of, near infrared reflectance spectroscopy applications in soil analysis: a review. Can. J. Soil Sci. 89, 531–541. https://doi.org/10.4141/CJSS08076.

Ng, W., Minasny, B., Jeon, S.H., McBratney, A., 2022. Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions. Soil Secur. 6, 100043. https://doi.org/10.1016/j.soisec.2022.100043.

Niazi, N.K., Singh, B., Minasny, B., 2015. Mid-infrared spectroscopy and partial least-squares regression to estimate soil arsenic at a highly variable arsenic-contaminated site. Int. J. Environ. Sci. Technol. 12, 1965–1974. https://doi.org/10.1007/s13762-014-0580-5.

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

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E. K., Vargas, R., Wetterlind, J., 2015. Soil spectroscopy: an alternative to wet chemistry for soil monitoring. Adv. Agron. 132, 139–159. https://doi.org/10.1016/bs.agron.2015.02.002.

Nriagu, J.O., Pacyna, J.M., 1988. Quantitative assessment of worldwide contamination of air, water and soils by trace metals. Nature 333, 134–139. https://doi.org/10.1038/333134a0.

O'Rourke, S.M., Minasny, B., Holden, N.M., McBratney, A.B., 2016. Synergistic use of Vis-NIR, MIR, and XRF spectroscopy for the determination of soil geochemistry. Soil Sci. Soc. Am. J. 80, 888–899. https://doi.org/10.2136/sssaj2015.10.0361.

Padilla, J.T., Hormes, J., Selim, H.M., 2019. Use of portable XRF: effect of thickness and antecedent moisture of soils on measured concentration of trace elements. Geoderma 337, 143–149. https://doi.org/10.1016/j.geoderma.2018.09.022.

Pāmu, 2025. In: Pāmu Farms of New Zealand, Landcorp Farming Limited (Ed.), Molesworth Station: New Zealand's Largest Farm. https://www.pamunewzealand.com/our-farms/molesworth-farm.

Pittaki-Chrysodonta, Z., Hartemink, A.E., Sanderman, J., Ge, Y.F., Huang, J.Y., 2021. Evaluating three calibration transfer methods for predictions of soil properties using mid-infrared spectroscopy. Soil Sci. Soc. Am. J. 85, 501–519. https://doi.org/10.1002/saj2.20225.

Ramirez Lopez, L., Stevens, A., Viscarra Rossel, R.A., Lobsey, C., Wadoux, A.M.J.C., Breure, T.S., 2016. Resemble: memory based learning in spectral chemometrics. R package version 2.0.0.

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

Rattenbury, M., Martin, A., Baisden, T., Turnbull, R., Rogers, K., 2018. Geochemical baseline soil surveys for understanding element and isotope variation across New Zealand. N. Z. J. Agric. Res. 61, 347–357. https://doi.org/10.1080/00288233.2018.1426616.

Rouillon, M., Taylor, M.P., 2016. Can field portable X-ray fluorescence (pXRF) produce high quality data for application in environmental contamination research? Environ. Pollut. 214, 255–264. https://doi.org/10.1016/j.envpol.2016.03.055.

RStudio Team, 2021. RStudio: Integrated Development Environment for R. RStudio. PBC, Boston, USA. RStudio version: 1.4.1103. http://www.rstudio.com/.

Sanderman, J., Baldock, J.A., Dangal, S.R.S., Ludwig, S., Potter, S., Rivard, C., Savage, K., 2021. Soil organic carbon fractions in the Great Plains of the United States: an application of mid-infrared spectroscopy. Biogeochemistry 156, 97–114. https://doi.org/10.1007/s10533-021-00755-1.

Sankey, J.B., Brown, D.J., Bernard, M.L., Lawrence, R.L., 2008. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. Geoderma 148, 149–158. https://doi.org/10.1016/j.geoderma.2008.09.019.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047.

Seidel, M., Hutengs, C., Ludwig, B., Thiele-Bruhn, S., Vohland, M., 2019. Strategies for the efficient estimation of soil organic carbon at the field scale with Vis-NIR spectroscopy: spectral libraries and spiking vs. local calibrations. Geoderma 354, 113856. https://doi.org/10.1016/j.geoderma.2019.07.014.

Shen, Z., Ramirez-Lopez, L., Behrens, T., Cui, L., Zhang, M., Walden, L., Wetterlind, J., Shi, Z., Sudduth, K.A., Baumann, P., Song, Y., Catambay, K., Viscarra Rossel, R.A., 2022. Deep transfer learning of global spectra for local soil carbon monitoring. ISPRS J. Photogrammetry Remote Sens. 188, 190–200. https://doi.org/10.1016/j.isprsjprs.2022.04.009.

Shenk, J.S., Westerhaus, M.O., Berzaghi, P., 1997. Investigation of a LOCAL calibration procedure for near infrared instruments. J. Near Infrared Spectrosc. 5, 223–232. https://doi.org/10.1255/jnirs.115.

Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66, 988–998. https://doi.org/10.2136/sssaj2002.9880.

Shrestha, G., Calvelo-Pereira, R., Roudier, P., Martin, A.P., Turnbull, R.E., Kereszturi, G., Jeyakumar, P., Anderson, C.W.N., 2022. Quantification of multiple soil trace elements by combining portable X-ray fluorescence and reflectance spectroscopy. Geoderma 409, 115649. https://doi.org/10.1016/j.geoderma.2021.115649.

Shrestha, G., Calvelo-Pereira, R., Poggio, M., Jeyakumar, P., Roudier, P., Kereszturi, G., Anderson, C.W.N., 2024. Predicting cadmium fractions in agricultural soils using proximal sensing techniques. Ennviron. Pollut. 349, 123889. https://doi.org/10.1016/j.envpol.2024.123889.

Siebielec, G., McCarty, G.W., Stuczynski, T.I., Reeves III, J.B., 2004. Near- and mid-infrared diffuse reflectance spectroscopy for measuring soil metal content. J. Environ. Qual. 33, 2056–2069. https://doi.org/10.2134/jeq2004.2056.

Sila, A.M., Shepherd, K.D., Pokhariyal, G.P., 2016. Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties. Chemom. Intell. Lab. Syst. 153, 92–105. https://doi.org/10.1016/j.chemolab.2016.02.013.

Soriano-Disla, J.M., Janik, L.J., Rossel, R.A.V., Macdonald, L.M., McLaughlin, M.J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49, 139–186. https://doi.org/10.1080/05704928.2013.811081.

Stafford, A.D., 2017. Distribution of cadmium in long-term dairy soils, its accumulation in selected plant species, and the implications for management and mitigation. School of Agriculture and Environment. Doctor of Philosophy (PhD). Soil Science. Massey University, Manawatu Campus, New Zealand, p. 324. http://hdl.handle.net/10179/12980.

Stahlmann-Brown, P., 2023. Survey of rural decision makers. Manaaki Whenua - Landcare Research, New Zealand. doi:10.7931/9kdt-0g35.

Stenberg, B., Rossel, R.A.V., 2010. Diffuse reflectance spectroscopy for high-resolution soil sensing. In: Viscarra Rossel, R.A., McBratney, A.B., Minasny, B. (Eds.), Proximal Soil Sensing. Springer Netherlands, Dordrecht, pp. 29–47. https://doi.org/10.1007/978-90-481-8859-8_3.

Summerauer, L., Baumann, P., Ramirez-Lopez, L., Barthel, M., Bauters, M., Bukombe, B., Reichenbach, M., Boeckx, P., Kearsley, E., Oost, K., Vanlauwe, B., Chiragaga, D., Heri-Kazi, A., Moonen, P., Sila, A., Shepherd, K., Mujinya, B.B., Van Ranst, E., Baert, G., Six, J., 2021. The central African soil spectral library: a new soil infrared repository and a geographical prediction analysis. SOIL 7, 693–715. https://doi.org/10.5194/soil-7-693-2021.

Tóth, G., Hermann, T., Da Silva, M.R., Montanarella, L., 2016. Heavy metals in agricultural soils of the European Union with implications for food safety. Environ. Int. 88, 299–309. https://doi.org/10.1016/j.envint.2015.12.017.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth Sci. Rev. 155, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012.

Viscarra Rossel, R.A., Lee, J., Behrens, T., Luo, Z., Baldock, J., Richards, A., 2019. Continental-scale soil carbon composition and vulnerability modulated by regional environmental controls. Nat. Geosci. 12, 547–552. https://doi.org/10.1038/s41561-019-0373-z.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Ge, Y., Gomez, C., Guerrero, C., Peng, Y., Ramirez-Lopez, L., Shi, Z., Stenberg, B., Webster, R., Winowiecki, L., Shen, Z., 2022. Diffuse reflectance spectroscopy for estimating soil properties: a technology for the 21st century. Eur. J. Soil Sci. 73. https://doi.org/10.1111/ejss.13271.

Viscarra Rossel, R.A., Shen, Z., Ramirez Lopez, L., Behrens, T., Shi, Z., Wetterlind, J., Sudduth, K.A., Stenberg, B., Guerrero, C., Gholizadeh, A., Ben-Dor, E., St Luce, M., Orellano, C., 2024. An imperative for soil spectroscopic modelling is to think global but fit local with transfer learning. Earth Sci. Rev. 254, 104797. https://doi.org/10.1016/j.earscirev.2024.104797.

Wang, C., Li, W., Guo, M., Ji, J., 2017. Ecological risk assessment on heavy metals in soils: use of soil diffuse reflectance mid-infrared Fourier-transform spectroscopy. Sci. Rep. 7, 40709. https://doi.org/10.1038/srep40709.

Wang, S., Guan, K., Zhang, C., Lee, D., Margenot, A.J., Ge, Y., Peng, J., Zhou, W., Zhou, Q., Huang, Y., 2022. Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: assessing potential of airborne and spaceborne optical soil sensing. Remote Sens. Environ. 271, 112914. https://doi.org/10.1016/j.rse.2022.112914.

Weindorf, D.C., Chakraborty, S., 2020. Portable X-ray fluorescence spectrometry analysis of soils. Soil Sci. Soc. Am. J. 84, 1384–1392. https://doi.org/10.1002/saj2.20151.

Wetterlind, J., Stenberg, B., 2010. Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. Eur. J. Soil Sci. 61, 823–843. https://doi.org/10.1111/j.1365-2389.2010.01283.x.

Wu, C.Y., Jacobson, A.R., Laba, M., Kim, B., Baveye, P.C., 2010. Surrogate correlations and near-infrared diffuse reflectance sensing of trace metal content in soils. Water Air Soil Pollut. 209, 377–390. https://doi.org/10.1007/s11270-009-0206-6.

Zhang, X., Sun, W., Cen, Y., Zhang, L., Wang, N., 2019. Predicting cadmium concentration in soils using laboratory and field reflectance spectroscopy. Sci. Total Environ. 650, 321–334. https://doi.org/10.1016/j.scitotenv.2018.08.442.