Amplifying the power of proximal sensing techniques to assess the cadmium concentration in agricultural soils

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Soil Science

at Massey University, Palmerston North, New Zealand

Gautam Shrestha

2023

Abstract

Cadmium (Cd) accumulation in agricultural soils due to long-term phosphate fertiliser applications has raised concerns in New Zealand and globally due to the potential toxicity of Cd in food products. Elevated soil Cd concentration can enhance Cd availability for plant uptake, increasing the risk of food chain transfer. Cadmium management is generally achieved through reference laboratory methods to estimate Cd concentration in soil and plant samples. Reference laboratory methods of Cd analysis are precise; however, sample preparation and associated resource cost make them expensive. As a complementary method, proximal sensing techniques including visible-near-infrared (vis-NIR: 350–2500 nm), mid-infrared (MIR: 4000–400 cm^{-1}) reflectance and portable X-ray fluorescence (pXRF: 0–40 keV) spectroscopy have been successfully used to monitor elevated Cd levels in mining areas and in plants showing stress or toxicity symptoms due to Cd. However, application of such technologies in agricultural soils with low Cd concentration are relatively understudied. Hence, this study was conducted to amplify the power of three proximal sensing techniques to quantify Cd in soil samples from diverse soil orders, climatic conditions, land uses, and vegetations and plant samples for cost-effective Cd monitoring at regional to farm scale.

In this doctoral study, soil and plant samples were scanned using vis-NIR, MIR, and pXRF sensors. Topsoil samples were obtained from (1) the Otago-Southland regional survey (n=622), (2) a pastoral farm survey (n=87) including dairy and sheep and beef farms with long-term phosphate fertiliser application history, and (3) two independent glasshouse experiments using Pallic and Allophanic soils amended with increasing soil Cd concentrations, and with or without a model forage herb, chicory (*Cichorium intybus* L.). In both experiments, chicory aboveground biomass and root samples were scanned using the three sensors, along with a periodic collection of vis-NIR spectra from soil and plant *in-situ*. Total Cd was determined in all samples, while the distribution of Cd among geochemical fractions was studied in the pastoral farm survey samples only. Reference laboratory results and spectral information were combined to develop models for accurate Cd predictions.

For regional survey samples (n=622, 0.01–0.56 mg Cd/kg) including agricultural soils (47%), validation (v) results (n=124, 0.01–0.43 mg Cd/kg) showed Granger-Ramanathan

model averaging of outputs from models using individual pXRF, vis-NIR, and MIR data as input for partial least squares (PLS) – support vector machine regression performed optimally to quantify total soil Cd with normalised root mean square error ($nRMSE_v$) of 37% and concordance correlation coefficient (CCC_v) of 0.84. For agricultural soils (n=84, 0.10–1.20 mg Cd/kg), cross-validation (cv) results of models using individual vis-NIR, MIR, and pXRF data as input for PLS performed with $nRMSE_{cv}$ of 26%, 30%, and 31% and CCC_{cv} of 0.85, 0.77, and 0.75 respectively. For acid soluble (0.01–0.27 mg Cd/kg) and organic matter bound (0.02–0.27 mg Cd/kg) Cd, models using vis-NIR data performed with $nRMSE_{cv}$ of 11% and 33% and CCC_{cv} of 0.97 and 0.84, respectively. For exchangeable (0.003–0.25 mg Cd/kg) Cd, a model using MIR data as input performed with $nRMSE_{cv}$ of 40% and CCC_{cv} of 0.57. Using the Otago and Southland regional survey soil samples spectra as a soil spectral library (SSL), Cd concentration in the local set (agricultural soil samples) were quantified. A model using MIR data from the regional SSL pastoral soil subset (n=283, 0.01–1.31 mg Cd/kg) spiked with selected local set samples (n=12) with weights (×4) as input for LOCAL algorithm quantified local soil Cd with nRMSE of 38% and CCC of 0.78.

In the glasshouse experiments, Cd translocation factor (TF) values for chicory were calculated using proximal sensor data and the results showed a significant relationship (R^2=0.74, $p<0.001$) between measured and predicted TF values. A model using *in-situ* leaf clip vis-NIR spectra showed optimal performance to assess Cd concentration in aboveground chicory biomass with $nRMSE_{cv}$ of 28% and CCC_{cv} of 0.93. Among vegetation indices calculated 'blue green index 2' showed a significant ($p<0.01$) R^2 value (0.19, 0.36) in both experiments. Models using pXRF spectra as input showed optimal performance to predict chicory root (n=28, 0.86–25.79 mg Cd/kg) and Allophanic soil (n=112, 0.41–4.81 mg Cd/kg) Cd with $nRMSE_{cv}$ of 16% and 9% and CCC_{cv} of 0.95 and 0.99, respectively. A model using laboratory vis-NIR spectra showed optimal performance to quantify Pallic soil Cd (n=336; 0.17–5.45 mg Cd/kg) with $nRMSE_{cv}$ of 22% and CCC_{cv} of 0.97.

Optimal prediction models using proximal sensor data can potentially be used for rapid cost-effective analysis of Cd concentration in soil and plant samples. Quantitative models for soil Cd using a combination of complementary proximal sensors data and chemometrics could feasibly be deployed for long-term monitoring of soil Cd at concentrations below pXRF detection limits and with reduced matrix interference from

organic matter when compared to the individual techniques alone. The use of proximal sensing techniques to determine total soil Cd concentration in New Zealand agricultural soils has the potential to improve the scale and scope of long-term repeated monitoring of soil Cd concentration required under the framework of the national Tiered Fertiliser Management System. Reflectance spectroscopy could potentially be implemented to monitor plant-available and potentially-available soil Cd fractions to minimise plant Cd uptake. The use of a large soil spectral library to assess the local Cd concentration in agricultural soils could reduce the analytical cost to the farmers and allow intensive spatial and temporal monitoring of pastoral farms based on spectral analysis only. The use of *in-situ* and laboratory proximal sensor data to calculate bioconcentration and translocation factors could potentially support the evaluation of Cd food chain transfer risks. The spectral library developed from this doctoral study, including soil and plant root and aboveground biomass pXRF, vis-NIR, and MIR spectra with a wide range of Cd concentration can be used as reference materials for field level and airborne remote sensing studies.

Acknowledgements

My humble gratitude to supervisors Dr Roberto Calvelo-Pereira, Dr Pierre Roudier, Dr Gabor Kereszturi, Dr Paramsothy Jeyakumar, and Prof. Chris Anderson, for their insights, inputs, and guidance throughout this doctoral journey. My sincere thanks to Dr Matteo Poggio for guiding during the first year of PhD as a co-supervisor. My appreciations to Emeritus Prof. Mike Hedley for giving hope while I was desperate to get a PhD position.

I am grateful to Massey's doctoral scholarship for giving an opportunity to conduct a doctoral study in New Zealand. During PhD, I also got fund support from the Kathleen Spragg Agricultural research trust to participate in a training "Use of irradiation apparatus for the non-medical use of X-rays in security, inspection and analysis systems" (covering ORS code of practice C10) organised by GNS Science. I would like to thank School of Agriculture and Environment for paying fourth-year tuition fees. I am grateful to scholarships including Putea Tautoko Doctoral Financial Support Grant, and Helen E Akers Postgraduate Scholarship for financial support. I highly appreciate the support received from the Graduate Research School (GRS) team.

I would like to thank Dr Aaron David Stafford for soil samples, which were the starting point of doctoral journey. My appreciations to Dr Adam Martin and Dr Rose E Turnbull from GNS Science for entrusting with soil samples collected as a part of the national geochemical soil survey. In addition, their knowledge has given this PhD a wider perspective. I cannot forget the role Pierre played to connect me with Adam and Rose.

I am humbled to the feedback provided by doctoral examiners Dr Brendan Malone, CSIRO, Australia, Dr Reddy Pullanagari, Stoneleigh Consulting Limited, and Dr James Hanley, Massey University. I would like to acknowledge Reddy for his feedback on confirmation report and exam as well. I would like to acknowledge soil laboratory staffs: Ian Furkert and Bob Toes; plant growth unit staffs: Lesley Taylor, Dr Sunil Sapkota, and Su Liu; Manaaki Whenua- Landcare Research (MWLR), environmental chemistry laboratory staffs: Kishor Kumar Karakkattuillom-Sankarannampoothiri, Thilak Palmada, and Sujatha Senanayake for their technical support. I am grateful to Sharon Wright and Fiona Bardell for their administrative support. I would like to thank Dr Akinson Tumbure, Dr Nilusha Ubeynarayana, Dumsane Matse Themba, Dr Rupsa Chakraborty, Cecilia Rodrigo Gomez, Duy Tran, Dr Aung Myo Thant, and Dr Allan Siano for their assistance.

My thanks to Dr Diwas Khatri, Dr Sandeep Karna, and Dr Kamal Prasad Adhikari for kindly supporting during initial days in New Zealand. I am grateful to MWLR, especially Dr Andrew Manderson, Jeanette King, Ngaire Foster, Dr Ursulla Jewell, and Dr John Drewry, for giving opportunity to experience the professional work environment in New Zealand while finishing doctoral studies.

My sincere thanks to Dr David Guerena at Consultative Group for International Agricultural Research (CGIAR) for giving opportunity to participate in the training on soil spectroscopy at Soil-Plant Spectral diagnostic laboratory, World Agroforestry Centre conducted by Dr Keith Duncan Shepherd, Dr Andrew Sila and the team there. This training acted as a catalyst to obtain this doctoral position with scholarship. I would like to remember Himalayan College of Agricultural Sciences and Technology (HICAST), and Nepal Agricultural Research Council (NARC) for providing opportunity to develop career in soil science.

Finally, the most important contributors of my life: before, during, and after this doctoral study: Ishwori Shrestha (mother), Gauri Bhakta Shrestha (father), Deepa Lohala (wife), family, and friends whose day-to-day help, inspirations, and support led to the successful accomplishment of the doctoral journey.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Al	Aluminium
BCF	Bioconcentration factor
BMA	Bayesian model averaging
CCC	Concordance correlation coefficient
Cd	Cadmium
CEC	Cation exchange capacity
CMS	Cadmium Management Strategy
cv	Cross-validation
Fe	Iron
FTIR	Fourier transform infrared
GFAAS	Graphite furnace atomic absorption spectrometer
GRA	Granger-Ramanathan model averaging
LMU	Land management unit
MPAES	Micro plasma atomic emission spectrometer
MIR	Mid-infrared
Mn	Manganese
nRMSE	Normalised root mean square error
NZSC	New Zealand Soil Classification
PLS	Partial least squares regression
PLS-SVM	Partial least squares – support vector machine
pXRF	portable X-ray fluorescence
RMSE	Root mean square error
RPIQ	Ratio of performance to interquartile distance
R^2	Coefficient of determination
SOM	Soil organic matter
SSL	Soil spectral library
TE	Trace element
TF	Translocation factor
TFMS	Tiered fertiliser management system
vis-NIR	visible-near-infrared

.

Chapter 1

Introduction

1.1 Importance of Cd monitoring

Cadmium (Cd) is a soil trace element (TE) with no essential function in biological processes (Gratão et al., 2006). At natural low concentrations, Cd has no environmental and health impact (Hooda, 2010). However, soil Cd concentration has increased in many places worldwide as a function of both geogenic and anthropogenic activities (Nriagu and Pacyna, 1988). Geogenic pathways for increased Cd input into soils are hydrothermal activities and the weathering of Cd-rich minerals (Kabata-Pendias, 2010). Anthropogenic sources of Cd are phosphate fertiliser and pesticide applied to agricultural lands, waste disposal, and mining and refining activities (Martin et al., 2017; Turnbull et al., 2019). Elevated concentrations of soil Cd may enhance plant-available concentrations, leading to uptake and accumulation of this TE in edible plant tissues of economically important forage herbs such as chicory (*Cichorium intybus* L.) (Stafford et al., 2016). As a consequence, year-round pastoral grazing can expose animals (e.g., sheep and dairy cattle) to long-term dietary accumulation of Cd (Roberts et al., 1994). Such exposures can have adverse effects on animal-based food product quality, resulting in dietary and trade risks when provisional safe limits for Cd are exceeded (ATSDR, 2012; Clemens et al., 2013; McDowell and Gray, 2022). Long-term ingestion of Cd via food, drink, skin contact with soil and water with an elevated level of Cd, or tobacco smoking, has been related to human health problems including kidney and bone damage (e.g., *Itai-itai* disease in Japan) and increased cancer risks (Godt et al., 2006; Järup and Åkesson, 2009; Nogawa et al., 2004).

International and national regulations are in place to minimise Cd risks to health, trade, land use flexibility, and the environment (CCME, 1999; FAO/WHO, 2011; FSANZ, 2016; MIEF, 2007). The FAO and WHO joint committee has set the tolerable limit for Cd at 7 µg/kg body weight/week (FAO/WHO, 2011). The Food Standards Australia New Zealand has set the maximum level of 0.1 and 0.05 mg Cd/kg fresh weight in leafy vegetables and animal meat, respectively (FSANZ, 2016). Europe and Canada have

established soil guideline values of around 1.5 mg/kg soil (CCME, 1999; MIEF, 2007). In New Zealand, a National Cd Management Strategy (CMS) was developed in 2011 and then implemented a tiered fertiliser management system (TFMS) to minimise risks and manage Cd accumulation in agricultural soils (CWG, 2011). The TFMS sets out increasingly stringent restrictions on the choice and rate of phosphate fertiliser application as total soil Cd concentration increases above the background concentration of 0.6 mg Cd/kg soil (Fertiliser Association, 2019). In addition, regional soil quality monitoring programmes assess soil Cd to limit the accumulation of this TE in soil and its transfer to water sources (Waikato Regional Council, 2016).

Cadmium analysis is critical to the successful implementation of any Cd management strategy. Reference methods for total Cd analysis are based on sample preparation including dissolving Cd from the matrix (e.g., soil, plant) with strong acids before quantifying Cd in a liquid solution (Kovács et al., 2000). These wet chemistry methods have been developed with knowledge of the interaction and association of Cd with the solid matrix (Tessier et al., 1979). Sequential extraction methods are used to investigate the distribution of Cd and other TE in the soil solid phase (Tessier et al., 1979). Additional insight into soil Cd fractions may help to understand Cd mobility and plant availability (Gleyzes et al., 2002). Among Cd fractions, the amount of exchangeable Cd is generally used as a proxy for plant-available Cd, whereas acid soluble and organic matter bound Cd can become plant-available with soil properties such as an acidic soil pH, high cation exchange capacity, and the dissolution of Fe/Mn oxides (Adriano, 1986; Hooda, 2010; Naidu et al., 1994). Such information is essential to understand plant Cd uptake (Krishnamurti, 2008). In addition, plant Cd analysis can provide important information to evaluate food chain transfer risks from pastoral grazing systems (Loganathan et al., 1999).

Studies comparing the relationship between soil type (e.g., Allophanic and non-Allophanic soils) on the distribution and plant uptake of Cd in New Zealand agricultural soils have been conducted, but such studies have generally had limited scope (Cavanagh et al., 2019; Gray et al., 2000; Stafford et al., 2018c). Gray et al. (1998) used four soil orders and a total of six soil samples to compare the influence of soil pH and contact time on Cd sorption and desorption. Stafford et al. (2018c) compared total Cd concentration between two long-term (>50 years) dairy farms with contrasting soil types. Gray et al. (2000) used a total of 12 samples representing seven soil orders to study soil Cd fractions.

Salmanzadeh et al. (2016) studied Cd fractions in two soil types on a total dataset of 30 samples. Cavanagh et al. (2019) compared plant Cd uptake in two contrasting soil types with soil Cd concentrations below 1 mg Cd/kg soil. Literature therefore shows that soil types influence total soil Cd and Cd fractions, as well as plant uptake. A better understanding of soil Cd fractions and Cd accumulation in plants growing in soils of contrasting soil Cd concentrations can contribute to the effective management of this TE.

Wet chemistry methods for intensive Cd monitoring are associated with time-consuming sample processing, rigorous health and safety requirements as well as standardisation and repeatability protocols that lead to an increase in total analysis cost (Stafford et al., 2018b). To reduce the cost of wet chemistry laboratory analysis, soil Cd predictions using factors that control Cd availability including soil pH, soil organic matter content, and total Fe have been proposed (Reiser et al., 2014). Similarly, estimates of plant Cd concentration based on soil properties have been proposed (Anderson et al., 2022; Cavanagh et al., 2019; Yi et al., 2020). However, the accuracy of such linear regression models depends on soil type, land use, and climatic conditions, which are highly variable (Cavanagh, 2014; Gray and Cavanagh, 2022; Gray et al., 2019a). An alternative approach is the use of novel technologies such as proximal sensing techniques. Such technology could potentially be used to determine Cd concentrations in New Zealand agricultural soils, improving the scale and scope of long-term repeated monitoring of Cd concentrations required under the framework of the TFMS.

1.2 Cd assessment using proximal sensing techniques

Proximal sensors record the response of a matrix to electromagnetic radiation when the detector is in contact or proximity (Viscarra Rossel et al., 2011). Proximal sensing techniques are non-destructive and provide spectra of the matrix containing extensive information on the molecular and compositional chemistry, allowing analysis of multiple soil properties (Soriano-Disla et al., 2014). These techniques can acquire data rapidly and represent a cost-effective way to cover spatial and temporal variations required for data-driven decision-making processes (Shepherd et al., 2022). Proximal sensing using spectroscopy has been established as a reliable method to assess the physical, chemical, and biological properties of a variety of matrices including soil and plant (Brevik et al., 2016; Shepherd and Walsh, 2002). Proximal sensing techniques have been widely tested

to predict the Cd concentration in contaminated soils (Nawar et al., 2019). The reflectance spectroscopy techniques visible-near-infrared (vis-NIR: 350–2500 nm) and mid-infrared (MIR: 4000–400 cm^{-1}) assess soil Cd based on the association of this TE with spectrally active soil components including soil organic matter and Al–, Fe–, and Mn–containing minerals (Soriano-Disla et al., 2014; Wu et al., 2010). Soil vis-NIR spectra contain overtones and combinations of vibrations of molecular bonds and functional groups (Ben-Dor, 2002). Leaf reflectance spectra relate to pigment content, cell structure, and biochemical composition and changes, which could be related to the deposition of Cd within plant tissues (Bandaru et al., 2010; Kooistra et al., 2004; Rathod et al., 2015a; Sharma and Dubey, 2006). Soil MIR spectra correspond to the fundamental vibrations due to the presence of specific Al–, Fe–, and Mn–containing minerals and organic functional groups (Siebielec et al., 2004). Plant MIR spectra contain fingerprints of stretching and bending vibrations of fundamental organic bonds (Artz et al., 2008; Usman et al., 2019). The X-ray fluorescence spectroscopy technique (portable XRF: 0–40 keV) detects specific fluorescent photons released by elements including Cd within the analyte matrix (Kalnicky and Singhvi, 2001). Portable-XRF can directly quantify Cd in both soil and plant matrices (Rouillon and Taylor, 2016; Towett et al., 2015a). However, there are challenges associated with the use of pXRF to estimate multiple trace elements at low (background) concentrations (Padilla et al., 2019; Ravansari et al., 2020; Rouillon and Taylor, 2016).

Spectral libraries are being developed around the world as a basis for the wider application of these proximal sensing techniques for the accurate estimation of soil and plant properties at a local scale (Brown, 2007; Shepherd and Walsh, 2002). A spectral library uses both spectra and reference laboratory results to develop the knowledge basis for the relationship between matrix property and spectral response (Viscarra Rossel et al., 2016). Soil spectral libraries (SSLs) have been developed for vis-NIR and MIR sensors at contrasting scales: local, regional/national, continental, and global (Dematt̂e et al., 2019; Summerauer et al., 2021; Viscarra Rossel et al., 2016). In the context of New Zealand, vis-NIR and MIR SSLs are being developed by Manaaki Whenua – Landcare Research for soil properties including soil total carbon, moisture, and pH (Blaschek et al., 2019; Malone et al., 2018; Roudier et al., 2020).

Advances in chemometrics have amplified the application of proximal sensing techniques for data-driven soil research (Wadoux et al., 2021b). Spectral information (predictors) is

calibrated using reference laboratory results by applying chemometric methods to assess qualitative and quantitative properties of interest (Nocita et al., 2015). Partial least squares (PLS) regression is one of the most successfully used algorithms to analyse proximal sensor data to assess soil properties including Cd concentration (Nawar et al., 2019). In addition, for large datasets, advanced chemometric methods including data fusion, model fusion, model averaging, and memory based algorithms have been implemented to assess soil properties including Cd concentration (Blaschek et al., 2019; Hong et al., 2019; O'Rourke et al., 2016b; Xu et al., 2020). Data fusion combines the spectral response of two or more sensors to increase predictive accuracy (Viscarra Rossel et al., 2006). The PLS-support vector machine (PLS-SVM) model fusion approach combines features of PLS to perform matrix decomposition of sensor and concentration data to extract latent variables and SVM to consider non-linear relationship between the spectral response and concentration data (Blaschek et al., 2019). Model averaging methods such as Granger Ramanathan model averaging (GRA) combines model outputs based on individual sensor data into a single prediction model (O'Rourke et al., 2016b). Memory-based algorithms including the LOCAL algorithm use a single regression algorithm to develop prediction models specific to each sample (Shenk et al., 1997).

Studies reporting the applicability of proximal sensing techniques to large-scale soil Cd monitoring (O'Rourke et al., 2016a; Xu et al., 2020) and the assessment of low Cd concentration soils are limited (Stafford et al., 2018b; Wang et al., 2017). Stafford et al. (2018b) estimated total soil Cd concentrations at two dairy farms based on the significant relationship between total soil Cd and vis-NIR predicted soil total carbon and total nitrogen. Only vis-NIR spectroscopy has been used to assess soil Cd fractions (Cipullo et al., 2019). Studies applying proximal sensing techniques to estimate plant uptake of Cd from low Cd concentration soils are similarly limited. Visible-NIR has been used to quantify plant Cd only at high concentrations (up to 800 mg Cd/kg aboveground biomass) (Feng et al., 2019), whereas MIR has been used for assessing plant composition, plasticity, and disease (Largo-Gosens et al., 2014; Palacio et al., 2014; White et al., 2011; Zhang et al., 2017). Studies using pXRF to quantify TE concentration in plant samples are limited (McGladdery et al., 2018; McLaren et al., 2012).

Controlled experiments using specific soil types and known concentrations of Cd including plants can be used to identify important spectral regions by deriving PLS loadings or variable importance in projection (VIP) that may contribute to the assessment

of low Cd concentrations on agricultural farms (Farrés et al., 2015; Kuhn and Johnson, 2013; Rathod et al., 2018). Controlled experiments have been conducted to assess plant Cd accumulation at toxic concentrations using vis-NIR spectroscopy (Rathod et al., 2015a; Sridhar et al., 2007). A robust spectral library developed from a large-scale survey could improve the prediction of Cd concentrations across areas with different soil types and plants with a wide concentration range that is characteristic of farm-scale mapping. Moreover, pXRF spectroscopy technique is less used to assess Cd concentration in agricultural soils. The possibility of using a large-scale spectral library to quantify local Cd has been poorly explored (Ng et al., 2022a).

This thesis fills the knowledge and resource gap on the potential of using proximal sensing techniques including vis-NIR, MIR, and pXRF independently and in combination to quantify low Cd concentrations at regional and farm scales. Widespread use of proximal sensing techniques as an accurate, rapid, non-destructive, and cost-effective method to monitor Cd would inform management practices for this TE, to protect highly productive lands from losing their land use flexibility (CWG, 2011; Drewry et al., 2021; Ministry for the Environment and Ministry for Primary Industries, 2022).

1.3 Research objectives

This thesis aims to develop robust prediction models using proximal sensors data including vis-NIR, MIR, and pXRF independently or in combination allowing accurate assessment of low Cd concentrations from regional to farm scale. The specific research objectives are:

- To quantify Cd and other TE (As, Cr, Hg, Mn, Ni, Pb, and Zn) in samples collected from a previous geochemical baseline soil survey of the Otago and Southland regions of New Zealand.
- To predict total Cd and distribution of Cd in geochemical fractions in a local dataset of agricultural soil samples collected from a previous study.
- To determine the potential of customising prediction models using sensor data to predict local soil Cd concentrations (agricultural soils) while leveraging a large soil spectral library (Otago-Southland).
- To quantify soil and plant Cd in chicory plants growing in a controlled environments with two distinct soils treated with increased concentrations of Cd.

1.4 Thesis outline

This thesis consists of seven chapters, including an introduction (Chapter 1), a literature review (Chapter 2), presentation of research works undertaken (Chapters 3–6), and a presentation of key findings, implications, and opportunities for future work (Chapter 7).

Chapter 1 introduces the research topic focusing on the application of proximal sensing techniques to assess the Cd concentrations in soil and plant samples and defines the objectives of this doctoral study.

Chapter 2 reviews the knowledge on i) the status, sources, and factors affecting the movement of Cd and other TE in New Zealand soils, ii) wet chemistry methods for total Cd and Cd fractionation analysis, and iii) the principles and applications of proximal sensing techniques to assess Cd and other TE.

Chapter 3 deploys proximal sensing techniques in combination with chemometric methods including PLS, PLS-SVM model fusion, sensor data fusion, and GRA model averaging to quantify low concentrations of Cd and other TE (As, Cr, Cu, Hg, Ni, Pb, and Zn) in samples collected and analysed from a previous geochemical baseline survey in Otago and Southland regions of New Zealand.

Chapter 4 reports the results of work that aimed to predict concentrations of total Cd and Cd fractions using PLS in a local set including only agricultural soil samples collected from a previous study.

Chapter 5 describes a strategy to customise prediction models using proximal sensor data as input for PLS and LOCAL algorithms to quantify local soil Cd at a farm scale while leveraging a soil spectral library (Otago and Southland regions).

Chapter 6 analyses reflectance and fluorescence spectra as input for PLS to assess Cd concentrations in soil and plant samples from controlled glasshouse experiments using Pallic and Allophanic soils with elevated soil Cd concentrations and chicory as a model forage crop species.

Chapter 7 presents the key findings of the doctoral study with implications for Cd monitoring and management and outlines opportunities for future work.

Chapter 2

Literature review

2.1 Cadmium and other TE in New Zealand agricultural soils

2.1.1 Origin and occurrence

Understanding TE abundance and variation in natural ecosystems is important for environmental management, agricultural production, and for protecting human health (Keesstra et al., 2016; Shrestha et al., 2021; Stojsavljevic et al., 2021). Trace elements such as cadmium (Cd), copper (Cu), and zinc (Zn) occur naturally in soils at low concentrations that cause no environmental concern (Hooda, 2010). Some of these TE, such as Cu and Zn, are essential elements for living organisms, and deficiency can cause health problems (Marschner, 2012; Prasad, 2012). However, soil TE concentrations can increase to above background or normal levels because of both geogenic and anthropogenic activities (Fig. 2.1 example Cd; Martin et al. (2016); Nriagu and Pacyna (1988)). Common geogenic pathways for increasing TE input are geothermal activity, volcanic eruptions, weathering and soil erosion (Kabata-Pendias, 2010; Zanders, 1998), while anthropogenic activities include phosphate fertiliser and pesticide application to agricultural land, waste disposal, vehicle emissions, and mining and refining activities (Harvey et al., 2017; Kubier et al., 2019; Morgan, 2010). Elevated TE concentrations in soil may enhance TE phyto-availability, leading to accumulation in plants and transfer through the food chain, potentially causing environmental, economic, and health risks (Alloway, 2013). Multiple edaphic and management factors influence the availability and mobility of TE in soils (Fig. 2.1).

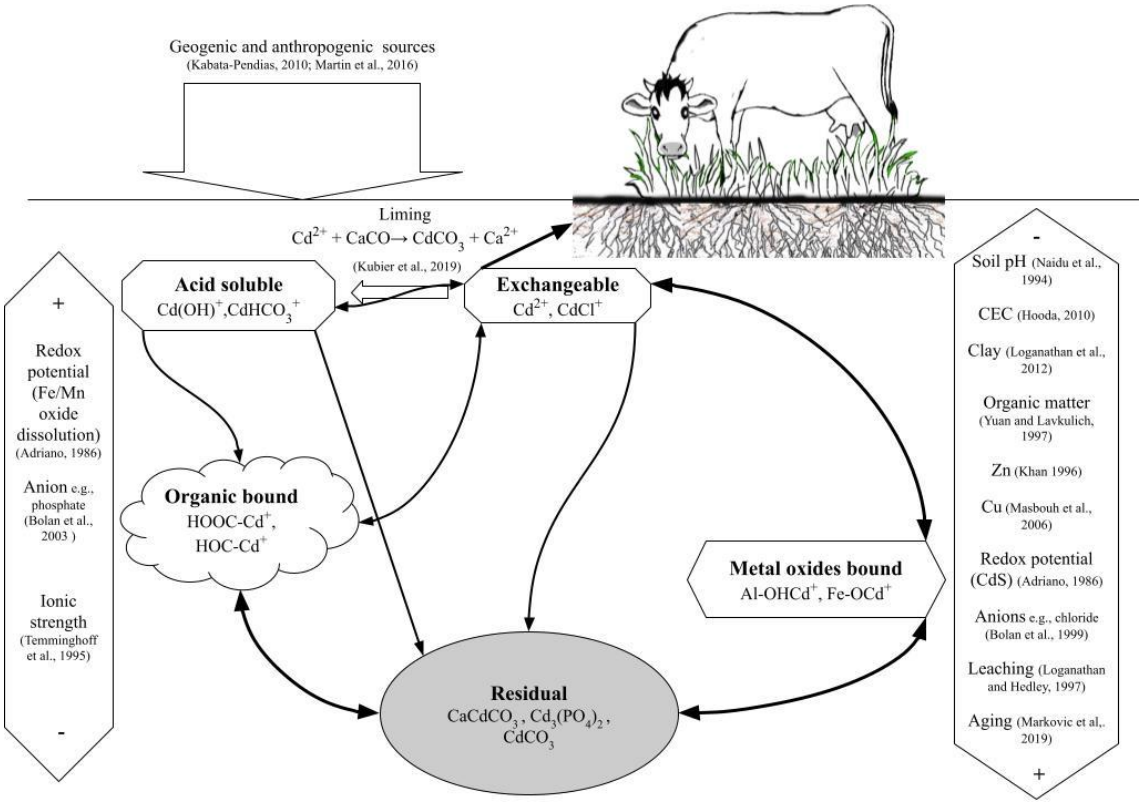


Fig. 2.1 Factors affecting mobility and accumulation of Cd in different soil fractions.

The concentration of Cu and Zn described in historical maps of New Zealand soils are categorised into <5 to >40 mg Cu/kg soil and <30 to >100 mg Zn/kg soil in the top 15 cm of the soil profile (Table 2.1; Soil Bureau (1962)). According to the New Zealand Soil Classification (NZSC) orders (Hewitt et al., 2021), Podzol soils show the lowest average (7 mg Cu/kg soil) total Cu concentration, whereas Allophanic soils show the highest average (27 mg Cu/kg soil) (Table 2.1). For Zn, the lowest average concentration is found in Pumice soils (39 mg Zn/kg soil) and the highest average in Granular soil (87 mg Zn/kg soil) (Table 2.1). Ultic soils show the lowest average Cd concentration (0.09 mg Cd/kg soil) whereas Allophanic soils often have the highest average concentration (0.74 mg Cd/kg soil) (Table 2.1). Allophanic and Pumice soils, formed from volcanic parent materials (Hewitt, 2010), contain more Cd than other soil orders (Table 2.1; Cavanagh (2014)). The highest reported Cd concentration is 2.93 mg Cd/kg soil for Allophanic soils under long-term dairy farming (Table 2.1; Abraham (2018)). The background Cd concentration in soil ranges between 0.03 and 0.14 mg Cd/kg soil, whereas Cd

concentration in pastoral soils ranges between 0.10 and 0.75 mg Cd/kg soil with an average value of 0.43 mg Cd/kg soil (Table 2.1; Rys et al. (2008)).

Table 2.1 Concentrations of Cd, Cu, and Zn (mg/kg soil) in New Zealand soils summarised as per historical data, the NZSC soil orders, and land uses.

Sources	Cd			Cu			Zn		
	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
	mg/kg soil								
Historical data[1]	–	–	–	< 5	–	> 40	< 30	–	> 100
NZSC soil orders[2]									
Allophanic	0.05	0.74	2.93	7.00	27.77	65.00	30.00	60.92	176.00
Brown	0.00	0.19	2.08	2.36	18.50	26.75	7.50	57.09	133.20
Gley	0.02	0.32	0.90	4.85	18.09	37.50	20.00	63.96	163.30
Granular	0.05	0.52	0.65	11.00	19.30	75.00	50.00	87.18	119.00
Melanic	0.10	0.25	0.56	6.61	23.32	74.65	33.70	76.10	148.60
Organic	0.03	0.42	1.52	7.50	21.34	60.00	17.50	42.29	96.00
Pallic	0.01	0.17	0.64	3.00	11.94	17.50	10.20	60.45	203.90
Podzol	0.01	0.11	0.25	4.45	7.39	48.52	26.30	51.65	105.20
Pumice	0.04	0.47	1.64	4.00	10.85	20.00	25.00	39.06	71.00
Recent	0.03	0.16	0.45	5.00	21.13	51.00	30.00	74.65	160.00
Ultic	0.06	0.09	0.67	2.50	9.84	41.00	9.00	40.71	110.00
Land use[3]									
Background	0.03	–	0.14	7.60	11.55	16.00	22.00	28.29	34.86
Pasture	0.10	0.43	0.75	11.30	13.52	16.70	50.70	59.51	66.60

[1](Soil Bureau, 1962)
[2](Cavanagh, 2014; Hewitt, 2010; Jeyakumar et al., 2010; Longhurst, 2006; Martin et al., 2017; Salmanzadeh et al., 2016; Stafford et al., 2018c; Taylor et al., 2007; Thompson-Morrison, 2017)
[3](Cavanagh et al., 2015; Martin et al., 2017; McDowell et al., 2013; Salmanzadeh et al., 2016; Taylor et al., 2010)

2.1.2 Accumulation of Cd in New Zealand agricultural soils and plants

New Zealand research published during the 1990s highlighted the existence of Cd concentrations above the background level in agricultural soils (Roberts et al., 1994; Zanders, 1998). Continuous application of phosphate fertiliser over the past 70 years for agricultural production containing traces of Cd (Abraham, 2018) has unintentionally augmented the Cd concentration in some agricultural soils (Fig. 2.1; Table 2.1). Compared to other farm management types, land subject to long-term intensive dairy production has recorded the highest Cd concentrations, up to 2.93 mg Cd/kg soil (Abraham, 2018).

In soil with elevated Cd concentration, a fraction of total soil Cd exists in a form that is available for plant uptake (Fig. 2.1; Gray et al. (1999b)). Consequently, plants can uptake and accumulate soil Cd in their biomass (Gray et al., 1999a). Many cultivated crops, such as cereals (Gray et al., 2019b), vegetables (Cavanagh et al., 2019; Gray et al., 2019a), and pasture herbs (Stafford et al., 2016), are known to accumulate Cd. Cadmium accumulation by different plant species decreases in the order of leafy vegetables>root vegetables>grain crops (Gray et al., 1999a). Amongst pasture herbs cultivated in New Zealand, chicory (*Cichorium intybus* L.) accumulated the greatest concentration of Cd: a comparative study by Stafford et al. (2016) of a range of pasture species grown in a greenhouse showed a Cd concentration in chicory of up to 1.64 mg Cd/kg dry matter while the dominant pasture crops ryegrass/white clover accumulated 0.06 to 0.12 mg Cd/kg DM (Stafford et al., 2016). Consequently, year-round grazing on pasture can potentially expose animals (e.g., sheep, dairy cattle) to long-term dietary accumulation of Cd (Roberts et al., 1994), especially where there is increased abundance of chicory in an animal's diet. Such exposure can have a potentially adverse effect on animal food product quality resulting in dietary (Godt et al., 2006) and trade risks (Kim, 2005) when the provisional safe limit for Cd uptake is exceeded (ATSDR, 2012; Clemens et al., 2013; McDowell and Gray, 2022).

2.1.3 Adverse health effects of long-term Cd ingestion

Exposure to Cd can happen through ingestion of food prepared from plants and animals exposed to elevated levels of total soil Cd, smoking tobacco, inhaling Cd from air polluted by vehicles and industrial emissions, and through working or playing in soil or water with elevated Cd levels (Godt et al., 2006). Long-term ingestion of Cd has been related to renal dysfunction, bone damage (e.g., *Itai-itai* disease in Japan), reproductive and gastro-intestinal disorders, respiratory and skin problems, as well as cancers (Järup and Åkesson, 2009; Nogawa et al., 2004). To minimise further Cd accumulation risks in New Zealand agricultural soils, a Cd management strategy is being actively pursued (CWG, 2011).

2.1.4 Regulation and management of soil Cd in New Zealand agriculture

The New Zealand government formed the national Cadmium Management Strategy (CMS) in 2011 to minimise Cd accumulation risk on health, trade, land use flexibility, and the environment over 100 years (CWG, 2011). A key component of the CMS is the Tiered Fertiliser Management System (TFMS) which aims to control the accumulation of Cd in agricultural soil by regulating future phosphate fertiliser applications.

The TFMS defines five Cd tier levels based on soil Cd concentrations, each with a total Cd concentration trigger value as a guideline for phosphate fertiliser application: 0.6 mg Cd/kg soil (tier 1), 1.0 mg Cd/kg soil (tier 2), 1.4 mg Cd/kg soil (tier 3), and 1.8 mg Cd/kg soil (tier 4) (Table 2.2). At the lower tiers, the choice and rate of phosphate fertiliser application are flexible; however, constraints on phosphate fertiliser use (and as a result, Cd addition to soil) increase at the higher tiers. At a soil Cd concentration of 1.8 mg Cd/kg soil, a farmer is restricted from applying any phosphate fertiliser to land (Table 2.2).

Table 2.2 Phosphate fertiliser choice and rate prescribed by the TFMS based on total soil Cd concentrations (Abraham, 2018; Fertiliser Association, 2019).

Tiers	Total soil Cd (mg/kg)	Phosphate fertiliser (kg/ha/year)			Descriptions
		280*	220*	100*	
0	<0.6				• Total soil Cd concentration is within the natural background concentrations. • Phosphate fertiliser type and rate can be chosen.
1	0.6 to <1.0	45	60	60–80	• Implementation of an appropriate management strategy is required to minimise soil Cd accumulation.
2	1.0 to <1.2	35	45	45–80	• Moderate restriction on the application of phosphate fertiliser type and rate.
	1.2 to <1.4	25	30	30–75	
3	1.4 to <1.5	20	25	25–55	• High restriction on the application of phosphate fertiliser type and rate.
	1.5 to <1.6	15	20	20–45	
	1.6 to <1.7	12.5	15	15–40	
	1.7 to <1.8	7.5	10	10–25	
4	≥1.8				• No phosphate fertiliser application. • Detailed site-specific soil testing for total Cd concentration. • Change land use and/or ploughing to a depth of at least 30 cm.

*Choice and rate as per Cd loading (mg Cd/kg P) in phosphate fertiliser.

During the revision of TFMS in 2019, an additional stepped reduction in phosphate fertiliser application recommendation was introduced in tiers 2 and 3 to extend the time necessary to reach the upper soil Cd threshold (Table 2.2).

The TFMS is a voluntary system, and to participate in the TFMS, an average dairy farm of size 150 ha should be divided into three land management units (LMU); each LMU should include six monitoring paddocks (for example, Fig. 2.2; Fertiliser Association (2016)). One composite sample prepared from 15–20 topsoil (0–15 cm) cores should be taken from each monitoring paddock for total soil Cd concentration using an approved laboratory analysis method. Repeated total Cd testing in an accredited laboratory is recommended every five years to assess the long-term trend of total soil Cd concentration, which increases the long-term costs for farmers (Fertiliser Association, 2016).

Fig. 2.2 Sampling scheme representation of the Tiered Fertiliser Management System showing each composite soil sample (black dots) containing 15–20 cores from a transect (black line) of each six monitoring paddock (demarked areas) representing each three land management units (LMUs shown in white, blue and red colours) in an average dairy farm of 150 ha size (Fertiliser Association, 2019).

2.1.5 Factors affecting mobility and accumulation of Cd in soils

Cadmium exists in soil as either water-soluble forms or insoluble organo-metallic complexes (Hooda, 2010). In an aqueous solution, Cd generally occurs as a divalent cation (Cd^{2+}) which can bond with anions to form water-soluble complexes such as $CdCl^{+}$, $CdCl_2^0$, $CdCl_3^-$, or $CdSO_4^0$. Cadmium can also form soluble complexes with dissolved organic matter (Kubier et al., 2019). Under acidic conditions, species including $CdCl^{+}$, $CdCl_2^0$, $CdNO_3^{+}$, $Cd(SO_4)_2^{2-}$ are possible whereas under alkaline conditions species including $CdHCO_3^-$, $CdCO_3^0$, $Cd(CO_3)_2^{2-}$, $CdOH^{+}$, $Cd(OH)_2^0$, $Cd(OH)_3^-$, Cd_2OH^{3+} are dominant (Fig. 2.3b). In the solid phase, Cd can be sorbed onto mineral surface functional groups including Fe–OCd^{+}, Fe–$OHCd^{2+}$; form stable complexes including $(FeO)_2HCdPO_3H^0$; or precipitate as $CaCdCO_3$ and $Cd_3(PO_4)_2$ (Bolan et al., 2003b). Cadmium can also be sorbed to soil organic matter via interaction with functional groups (e.g., carboxylic and phenolic groups) to form organic ligand–Cd^{2+}, organic ligand–Cd^- chemical structures (Fig. 2.1; Loganathan et al. (2012)).

The movement of Cd from one fraction to another depends on multiple soil properties, such as soil pH, redox potential, ionic strength, soil organic matter, clay minerals, cation exchange capacity (CEC), metal oxides, anion exchange capacity, and liming (Fig. 2.1; Adriano (1986); Gray et al. (2019b); Naidu et al. (1994); Stafford et al. (2018c); Temminghoff et al. (1995); Yuan and Lavkulich (1997)). The exchangeable Cd fraction is generally used to assess the availability of Cd to plants (Fig. 2.1; Gleyzes et al. (2002)). The sorption and desorption of Cd control movement of the exchangeable Cd fraction to/from other soil fractions (Fig. 2.1; Loganathan et al. (2012)).

2.1.5.1 Soil pH, redox potential, and ionic strength

Soil solution pH and redox potential (E_h) are the most critical factors influencing Cd solubility in soil (Fig. 2.3a; Hooda (2010)). At soil pH below 6, there is an increase in the activity of acidic cations in soil solution (H^{+}, Fe^{3+}, Al^{3+}), and their positively charged hydroxides, which compete with Cd^{2+} cations for pH-dependent sites on the soil solid phase (e.g., electrostatic negatively charged sites of surfaces in clays and organic matter contributing to CEC), promoting the release of Cd species into solution (Fig. 2.1; Cottenie and Verloo (1984)). At high soil pH, Cd cations bind to the hydroxyl ion forming $CdOH^{+}$

(Naidu et al., 1994) or carbonates to form $CdHCO_3^+$ (Xian and Shokohifard, 1989), and can also adsorb onto pH-dependent negatively charged soil surfaces (Fig. 2.3b).

Cadmium tends to be more mobile in oxidising conditions than in reducing conditions (Fig. 2.3a; Förstner (1993)). In oxidising conditions (positive E_h), Fe (III) and Mn (III-IV) are present mainly as insoluble hydrated oxides and Cd is released from the metal oxides fraction to the exchangeable fraction (Fig. 2.1; Hooda (2010)). In contrast, under reducing conditions, the presence of sulphides in the soil can cause cadmium sulphide (CdS) to precipitate (Adriano, 1986). Under oxidising conditions, Cd has a lower sorption affinity to metal oxides than other TE (e.g., Cu, Zn, Co) and this leads to a relatively higher Cd concentration in the exchangeable fraction than other TE (Hooda, 2010).

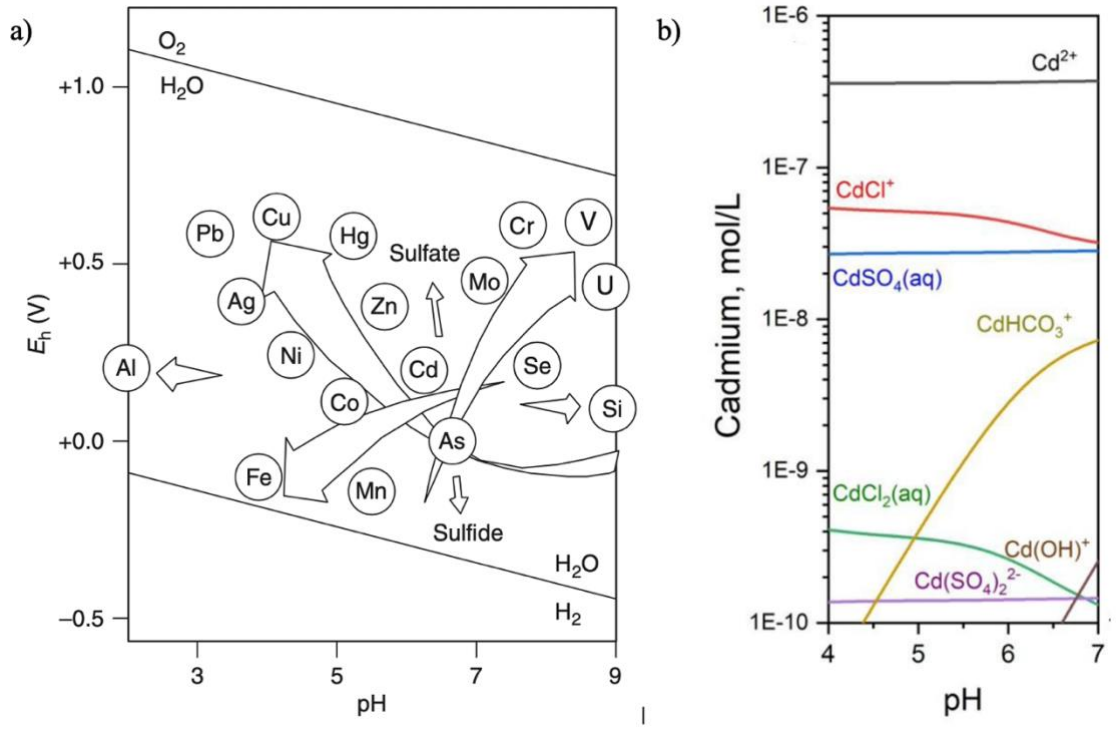


Fig. 2.3 a) Trace elements mobility as a function of redox potential and pH changes in solid waste material, broadening arrows represent increasing mobility; from Förstner (1993) © copyright#1993, reprinted by permission of Informa UK Limited, trading as Taylor & Francis Group, http://www.tandfonline.com and b) Concentration versus pH diagram showing Cd speciation from pH 4 to 7 for a solution with the composition of Cd ($10^{-6.4}$ M), Cl (10^{-3} M), ΣCO_2 (10^{-3} M) and SO_4 (10^{-3} M); Reprinted from Kubier et al. (2019), Copyright (2019), with permission from Elsevier.

Changes in the ionic strength of soil solution will often cause soil pH variation and hence changes in Cd sorption (Fig. 2.1; Salmanzadeh (2017)). Higher soil ionic strength means more cations compete for exchange sites (at clay surfaces or SOM), leaving Cd and other TE in the exchangeable fraction (Fig. 2.1; Temminghoff et al. (1995)). The ability of cations to compete for exchange sites with an increase in ionic strength is dependent upon the size-to-charge ratio of the cations (e.g., $Al^{3+}>Ca^{2+}>K^{+}>Na^{+}$; Wang et al. (2010); Zanders (1998)). Increasing ionic strength generally leads to the release of TE in the order: Cd>Zn>Cu (Spark et al., 1995) from sorption surfaces to the solution phase (Fig. 2.3a).

2.1.5.2 Clay minerals and soil organic matter

In soils, clay minerals are a significant source of negative surface charge (Hooda, 2010), influencing CEC. In the 2:1 layer silicate clays, characterised by a large permanent negative charge, Cd and other TE are electrostatically attracted (i.e., exchangeable) because of non-specific sorption (Loganathan et al., 2012). In 1:1 silicate clay, containing terminal Al–OH or Fe–OH functional groups, strong covalent bonds can be formed with cations including those of TE because of specific sorption (Fig. 2.1; Loganathan et al. (2012)). Poorly ordered alumino-silicate clay minerals (such as allophane and imogolite) contain high variable charge (pH-dependent) showing high specific sorption capacity (Christensen and Haung, 1999), thus influencing the retention of TE. Al, Fe and Mn oxides and hydroxides also represent variable charged minerals with strong specific sorption capacity for TE (Fig. 2.1; Christensen and Haung (1999)). The specific adsorption of Cd onto clay mineral and metal oxide surfaces ultimately leads to the transfer of available Cd to the so-called “residual soil Cd fraction” with time (Fig. 2.1; Kubier et al. (2019)).

Soil organic matter (SOM), at the pH of most soils in New Zealand, is negatively charged and can be a significant contributor to CEC (Hooda, 2010). Cadmium can be held by terminal functional groups such as carboxylic (−COOH) and phenolic (−COH) due to electrostatic attraction that results in non-specific sorption of the TE (Loganathan et al., 2012). The general order of affinity of different TE with SOM is: $Cu^{2+}>Cd^{2+}>Co^{2+}>Mn^{2+}>Zn^{2+}$ (Bolan et al., 2003b). Cadmium sorption by SOM at high pH is more stable due to the formation of organo-metallic complexes (Hooda, 2010). In

these conditions, specific sorption of Cd onto SOM surfaces is possible by forming covalent bonds, which are relatively strong, between Cd and multiple carboxyl and acid hydroxyl groups (Fig. 2.1; Young (2013)).

2.1.5.3 Other factors: anions in the soil solution, liming

Phosphate minerals, such as hydroxy-apatite, can bound Cd via surface complexation or co-precipitation where Cd^{2+} substitutes for Ca^{2+} at higher pH (Fig. 2.1; Bolan et al. (2003a)). Other anions such as chloride (Cl^-) decrease sorption by forming $CdCl^+$, which is more soluble than Cd^{2+} (Bolan et al., 1999). Liming adds free $CaCO_3$ to the soil, which reduces the solubility of Cd by forming Cd carbonate (i.e., an acid-soluble fraction) (Marković et al., 2019), which is only sparingly soluble or may precipitate, contributing to a residual form of Cd in soil (Fig. 2.1; Kubier et al. (2019)).

2.2 Cadmium analysis

Soil analysis is critical to any soil Cd management strategy based on total soil Cd concentration (Fig. 2.1; Table 2.3). Monitoring of soil Cd concentration under the TFMS typically relies on wet chemistry methods which are associated with a time-consuming process of sample preparation and analysis (Du Laing, 2010). Wet chemistry methods using sequential extractions can provide additional insight into Cd and other TE distributions among different soil components (e.g., metal oxides bound, organic matter bound). As an alternative to wet chemistry methods, several non-destructive techniques have been developed. These include instrumental neutron activation analysis, particle-induced X-ray emission, and X-ray absorption near edge structure spectroscopy. Those can allow direct measurement of Cd and other TE in soil samples (Du Laing, 2010; Siebers et al., 2013). However, these non-destructive techniques have a range of restrictions, such as limited sensitivity and limited portability, which can hinder the widespread use of these techniques (Du Laing, 2010), whereas proximal sensing techniques based on reflectance and fluorescence spectroscopy have proven useful in measuring multiple soil properties (Viscarra Rossel et al., 2006), including TE concentration (Nawar et al., 2019).

2.2.1 Total Cd concentration analysis

Reference laboratory methods for total TE (including Cd) concentration analysis use acids e.g., hydrofluoric acid (HF), hydrochloric acid (HCl), nitric acid (HNO_3), sulphuric acid (H_2SO_4), perchloric acid ($HClO_4$), including occasionally hydrogen peroxide (H_2O_2) alone or in combination to dissolve the solid matrix and release TE into solution (Du Laing, 2010). For example, HCl dissolves phosphates, carbonates, and some oxides and sulphides, whereas HNO_3 attacks matrices that are not dissolved by HCl (Hooda, 2010). The most widely used digestion acid mixture is aqua regia (1:3 HCl:HNO_3) (Salmanzadeh, 2017) due to its high oxidising ability. Hydrofluoric acid dissolves silica-based soil matrices (Hooda, 2010); however, its use in chemical analysis is limited due to its toxic nature (Zanders, 1998). Determination of Cd and other TE concentrations in digested aliquots can be made using inductively coupled plasma-atomic emission and mass spectrometry (ICP-AES and ICP-MS) (Kovács et al., 2000) and graphite furnace atomic absorption spectrometry (GFAAS) (Gray et al., 1999c).

2.2.2 Sequential extraction of soil Cd

The chemical forms of Cd and other TE in the soil solid phase can be studied using sequential extraction procedures, which recover elements from operationally defined soil fractions (Fig. 2.1; Adriano (1992)). Sequential extraction provides information about the origin, mode of occurrence, mobility, and biological availability of elements in the soil and such information is essential to understand plant uptake (Krishnamurti, 2008). When using a sequential extraction procedure, between three to ten steps are needed, applying extractants of increasing reactivity and strength. Sequential extraction methods broadly separate Cd into five geochemical fractions: exchangeable, acid soluble, metal oxides bound, organic matter bound, and residual fraction (Table 2.3; Gleyzes et al. (2002)). Among the sequential extraction methods described in the literature, that of Tessier et al. (1979) is the most common method in use (Du Laing, 2010).

Table 2.3 Fractions of Cd obtained following sequential extraction methods (Gleyzes et al., 2002; Tessier et al., 1979).

Fractions	Target phases	Mainly extracted geochemical forms
F1	Exchangeable	Weakly sorbed Cd (soluble and exchangeable)
F2	Acid soluble	pH sensitive, non-specifically sorbed Cd on the solid surface
F3	Metal oxides bound	Cd associated with amorphous or crystalline Fe, Mn, and Al oxides
F4	Organic matter bound	Cd sorbed to organic matter fraction
F5	Residual	Cd mainly associated with primary and secondary minerals in the crystalline lattice

In New Zealand, soil Cd fractions were studied using different sequential extraction procedures including the Tessier method (Table 2.4; Bolan et al. (2003a); Gray et al. (2000); Salmanzadeh et al. (2016)). Though some fractions are common in all three sequential extraction methods, organic matter bound fraction is at the second last of the

Table 2.4 Sequential extraction methods used for Cd fractionation of New Zealand soils.

Methods/Fractions	Extractants	Soil (g): Solution (ml)	Conditions
Tessier et al. (1979)			
Exchangeable	1.0 M $MgCl_2.6H_2O$ (pH 7)	1 in 8	Shaking 40 rpm 1 h
Acid soluble	1.0 M $CH_3COONa.3H_2O$ (pH 5)	1 in 8	Shaking 40 rpm 5 h
Metal oxides bound	0.04 M $NH_2OH.HCl$ in 25% (v/v) CH_3COOH	1 in 20	Refluxed 6 h at 96±3ºC
Organic matter bound	0.02 M HNO_3+30% H_2O_2 (pH 2)	1 in 3+5	Refluxed 2 h at 85±2ºC
	30% H_2O_2 (pH 2)	1 in 3	Refluxed 3 h at 85±2ºC
	3.2 M $NH_4CH_3CO_2$ in 20% HNO_3	1 in 5 diluted to 20	Shaking 40 rpm 30 min
Residual	concentrated HNO_3	1 in 10	Digestion 2 h at 120ºC
Sposito et al. (1982)			
Soluble + exchangeable	0.5 M KNO_3	1 in 12.5	Shaking 16 h
Organic-bound	0.5 M NaOH	1 in 20	Shaking 16 h
Inorganic-bound	0.05 M Na_2EDTA	1 in 25	Shaking 6 h
Residual	4 M HNO_3	1 in 25	Refluxed 16 h at 80ºC
Shuman (1985)			
Exchangeable	0.01 M $Ca(NO_3)_2$	5 in 30	Shaking 2 h
Organic-bound	3.5% NaOCl (pH 8.5)	5 in 10	Boiling water bath 30 min (repeat once)
Amorphous Fe oxides bound	0.2 M $(NH_4)_2C_2O_4.H_2O$	2 in 25	Shaking 4 h in the dark
Crystalline Fe oxide bound	0.2 M $(NH_4)_2C_2O_4.H_2O$+0.2 M $H_2C_2O_4$+0.1 M ascorbic acid (pH 3)	2 in 25	Boiling water bath 30 min
Residual	concentrated HNO_3	0.5 in 10	Microwave digestion

sequence in Tessier et al. (1979) procedure whereas in the other two protocols it is extracted after the exchangeable fraction (Table 2.4). Operationally defined fractions and chemicals used for extracting individual fractions differed as per extraction methods, whereas concentrated HNO_3 was commonly used to extract residual fraction (Table 2.4).

Recovery after sequential extraction ranges from 80–116% of total soil Cd (Table 2.5). Salmanzadeh et al. (2016) found concentrations of 0.2 mg Cd/kg in the exchangeable fraction, 0.05 mg Cd/kg in the acid soluble fraction, 0.25 mg Cd/kg in the metal-oxides bound fraction, 0.17 mg Cd/kg in the organic matter bound fraction, and 0.04 mg Cd/kg in the residual fraction (modified Tessier method) for a Waikato Allophanic soil under long-term pastoral land use (total Cd 0.8 mg/kg) (Table 2.5). Gray et al. (2000) found that only 3% of total Cd was recovered from the exchangeable fraction for both Allophanic and non-Allophanic soil following the extraction method by Shuman (1985), whereas Bolan et al. (2003a) quantified 34% of total Cd in the exchangeable fraction following the extraction method described by Sposito et al. (1982) (Table 2.5).

Table 2.5 Recovery percentage of total soil Cd in each fraction from sequential extraction studies of New Zealand soils.

Soil type/References	N	Total Cd (mg/kg)	Recovery % of total	Total recovery in each fraction (%)				
				Exchangeable	Acid soluble	Metal oxides bound	Organic matter bound	Residual
Allophanic soil								
Salmanzadeh et al. (2016)	30	0.80	89	28	7	35	24	6
Gray et al. (2000)	12	1.04	116	3	*	17	38	42
Bolan et al. (2003a)	2	0.3	109	34	*	37	17	12
Non-Allophanic soil								
Salmanzadeh et al. (2016)	30	0.78	87	29	7	35	22	6
Gray et al. (2000)	12	0.23	116	3	*	25	34	38

*Acid soluble fraction is not separately defined in these studies.

2.3 Cadmium assessment using proximal sensing techniques

Proximal sensing using spectroscopy has been evolving as a reliable soil testing method for quantitative and qualitative assessment of multiple physical, chemical, and biological properties in a fast and highly reproducible way (Brevik et al., 2016; Poggio et al., 2018; Shepherd and Walsh, 2002; Vasques et al., 2014). Proximal sensors are being widely deployed in a range of portable and handheld spectrometers (Nocita et al., 2015). Proximal sensors record the response of a matrix (e.g., soil, plant) to electromagnetic radiation (Viscarra Rossel et al., 2011; Xu et al., 2020). Spectral information (predictors) is calibrated using reference laboratory results (target property) by applying chemometric methods to assess qualitative and quantitative properties of interest (Nocita et al., 2015). Portable X-ray fluorescence (pXRF) spectroscopy quantifies total TE concentration directly, using built-in calibration models using a limited range of TE concentrations and spectral response of standard matrices (Lemière, 2018). However, TE detection using reflectance spectroscopy (e.g., visible-near-infrared (vis-NIR) or mid-infrared (MIR) spectroscopy) is based on the spectral response of spectrally active soil components (e.g., soil organic matter and Al–, Fe–, and Mn–containing minerals) with which TE are associated (Siebielec et al., 2004). This is because TE are spectrally "featureless" in reflectance spectra (Kooistra et al., 2001; Wu et al., 2010; Wu et al., 2007; Xia et al., 2007). Literature describing the prediction of multiple soil properties using spectral data is extensive (Nawar et al., 2019; Rossel and Webster, 2012). For Cd and other TE assessment in agricultural soil and plant samples, pXRF, vis-NIR, and MIR spectroscopy has shown a potential for wider applications (Fig. 2.4).

2.3.1 Portable X-ray fluorescence spectroscopy

Portable-XRF is a handheld sensor to capture spectral information across the 0–40 keV (0.001–10 nm) range of the electromagnetic spectrum. This sensor detects specific fluorescent photons released by each element within the analyte matrix, which enables quantitative and qualitative assessment of elements from sodium (Na) to uranium (U) (Kalnicky and Singhvi, 2001). Element-specific energy bands (K_{α}) enable elemental analysis using pXRF sensors, for example, 6.40 keV for Fe and 23.17 keV for Cd (Fig. 2.4; Table 2.6; Elam et al. (2002)).

Portable-XRF spectroscopy has become a well-established technique for screening elements in soils, including As, Pb, and Hg (Tables 2.6 and 2.7; Lemière (2018)). This spectroscopy technique quantifies elemental concentration based on either a built-in calibration model developed from X-ray response data of a limited matrix concentration range, or an analysis of raw spectra with chemometric methods for a local concentration range (O'Rourke et al., 2016a). However, the application of pXRF spectroscopy is constrained by detection limits (Table 2.7) which rarely approach those of wet chemistry-based methods, as well as matrix interference due to moisture, soil organic matter, and particle size (Padilla et al., 2019; Ravansari et al., 2020). There are no reported studies quantifying Cd fractions using pXRF. Studies using pXRF to quantify TE in plant samples are limited (McGladdery et al., 2018; McLaren et al., 2012; Towett et al., 2015a) and this technique has not yet been rigorously tested to quantify plant Cd concentrations.

Table 2.6 X-ray absorption and emission energies of TE and spectrally active soil components (Elam et al., 2002).

Elements	LOD (ppm)*	K edge	$K_{\alpha 1}$	$K_{\beta 1}$	L_3 edge	$L_{\alpha 1}$	$L_{\beta 2}$
		Energy value (keV)					
Trace elements							
As	2	11.867	10.543	11.726			
Cd	5	26.711	23.173	26.093			
Cr	15	5.939	5.900	6.492			
Cu	5	8.979	8.046	8.904			
Hg	1	83.102	70.818	80.255	12.284	9.989	11.906
Ni	5	8.333	7.480	8.267			
Pb	3	88.005	74.970	84.939	13.419	10.839	12.955
Zn	3	9.659	8.637	9.570			
Spectrally active soil components							
Al	650	1.559	1.486	1.557			
C[#]	–	0.284	0.277	–			
Fe	7	7.112	6.405	7.059			

*LOD=Limit of detection is the lowest quantity of an element in parts per million (ppm) that built-in calibration in portable X-ray fluorescence spectroscopy (pXRF) Olympus Vanta instrument with rhodium anode can detect in an interference-free silica blank.
[#]No mention of LOD value for carbon.

2.3.2 Visible-near-infrared spectroscopy

A vis-NIR spectrometer uses a sensor to record the reflected and emitted light from the matrix in the 350–2500 nm (25000–4000 cm^{-1}) region of the electromagnetic spectrum

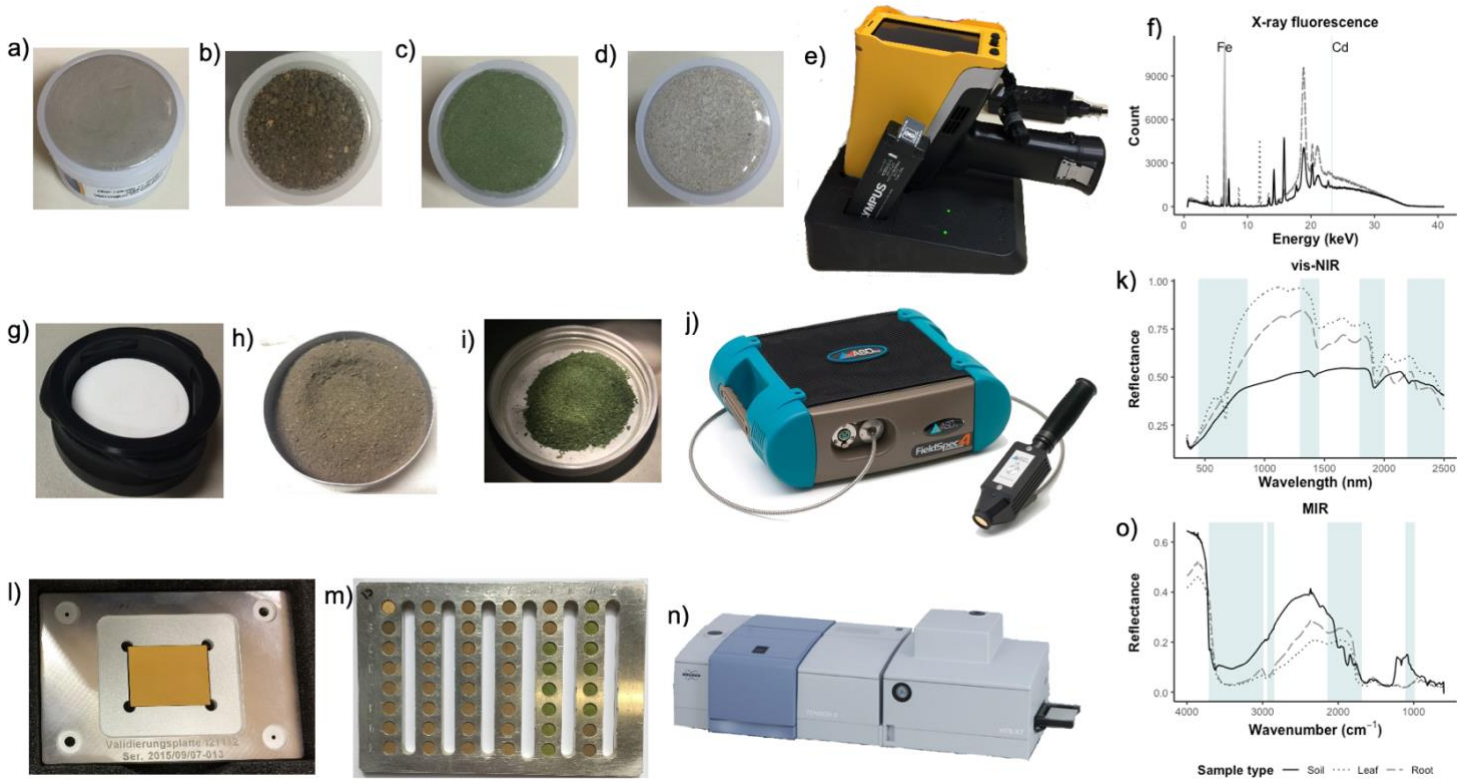


Fig. 2.4 Laboratory scanning by pXRF sensor: a) SRM2711a reference, b) soil, c) plant aboveground biomass, and d) plant root samples, e) Olympus VCR pXRF instrument, f) pXRF spectra, vis-NIR sensor: g) white reference Spectralon, h) soil, and i) plant aboveground biomass samples in the sample holders, j) ASD FieldSpec4 spectroradiometer, k) vis-NIR reflectance spectra, MIR sensor: l) validation plate, m) soil and plant samples in 48 wells microplate, n) FTIR spectrometer with HTS-XT (Tensor II), and o) MIR reflectance spectra.

(Fig. 2.4). Soil vis-NIR spectra capture overtones and combinations of vibrational molecular bonds and functional groups (Table 2.7). Among spectroscopy techniques, vis-NIR spectroscopy is the most widely used for soil TE estimation due to its broad applicability to quantify multiple soil properties, ease of sample preparation, and instrument cost (Table 2.7; Nawar et al. (2019)).

Soriano-Disla et al. (2014) described the application of vis-NIR spectra to quantify soil components across several ranges of the electromagnetic spectrum that can be associated with soil Cd: 400–850 nm (soil colour, Fe oxides, and soil organic matter), 1300–1450 nm (Al–, Fe–, and Mn–containing minerals), 1800–2000 nm (Al–, Fe–, and Mn–containing minerals and soil organic matter), and 2200–2500 nm (Mg–, Al–, Fe–, and Mn–containing minerals and soil organic matter) (Appendix 2.1). *In-situ* vis-NIR spectra have been used to assess plant growth, stress, and nutritional status by calculating vegetation indices (Appendix 2.2). A vegetation index is a unitless ratio or linear combination of two or more selected spectral bands, which is used to maximise plant spectral signal and minimise background effect (Rathod et al., 2013). In a comparison of *in-situ*- versus lab-based vis-NIR sensors to quantify soil Cd, the field generated spectra had lower predictive accuracy than laboratory-generated spectra due to changes in the particle size, moisture content, and illumination conditions experienced in the field (Zhang et al., 2019). Visible-NIR spectroscopy has been used to assess the fractions of TE including Cd in contaminated soils including landfill sites and TE polluted areas (Chakraborty et al., 2017; Chodak et al., 2007; Cipullo et al., 2019). Both *in-situ* and laboratory vis-NIR spectroscopy has been used to quantify plant TE at high concentrations showing stress or toxic effect in the plant (Cao et al., 2021; Cui et al., 2021; Feng et al., 2019; Sridhar et al., 2007) whereas possibility to use to quantify TE at low concentrations in plants deserves further study.

2.3.3 Mid-infrared spectroscopy

Mid-IR sensor records diffuse reflectance spectra across the 4000–400 cm^{-1} (2500–25000 nm) range of the electromagnetic spectrum and detects fundamental vibrations of molecular bonds and functional groups in an analyte matrix (Fig. 2.4; Table 2.7; O'Rourke et al. (2016a); Siebielec et al. (2004)). In soil MIR spectra, regions including the 3000–2700 cm^{-1} (minerals including gibbsite), 2929–2855 cm^{-1} (alkyl compounds), 2130–1700

cm^{-1} (metal–carbonyl bonds) and 1100–1000 cm^{-1} (quartz) are considered important for TE quantification (Appendix 2.3; Janik et al. (1998); Niazi et al. (2015); Wang et al. (2017)).

Soil MIR spectra exhibit typical peaks and shallows due to the presence of specific Al–, Fe–, and Mn–containing minerals, such as gibbsite (3528 cm^{-1} and 3461 cm^{-1}), kaolinite (3599 cm^{-1}), and quartz (1100–1000 cm^{-1}, 800 cm^{-1}); TE are commonly associated with these minerals in soil (Table 2.7; Appendix 2.3). Studies have used MIR sensors to quantify carbon fractions in the soil (Baldock et al., 2013), which shows potential to assess the soil Cd fractions as well. Mid-IR spectroscopy has also been used for the assessment of a range of characteristics of plant, including root composition, plasticity and disease (Largo-Gosens et al., 2014; White et al., 2011; Zhang et al., 2017).

2.4 Chemometrics for predictive modelling

Proximal sensors record light reflection and absorption across a defined wavelength range of the electromagnetic spectrum (e.g., X-ray, vis-NIR, MIR) when a detector is in contact or proximity (Viscarra Rossel et al., 2011). The sampling frequency is high, which often leads to a large number (i.e., hundreds to thousands) of wavelength readings. These correspond to n-dimensions, where *n* is the number of wavelengths. Individual wavelength values are highly correlated, requiring intense pre-processing via dimensionality reduction, or feature selection approaches (Ghamisi et al., 2017). Massart et al. (1997) describe chemometrics as a chemical discipline that uses mathematics, statistics, and formal logic to (a) design or select optimal experimental procedures, (b) provide maximum relevant chemical information by analysing chemical data, and (c) obtain knowledge about chemical systems. In the context of proximal sensing techniques, chemometric methods are used to uncover relationships within input data, both predictors (spectral data) and soil and plant property results (target) variables, to predict target property; this is described as predictive modelling (Fig. 2.5; Kuhn and Johnson (2013)).

In a supervised modelling approach, a training data set with known predictors and soil and plant property data are used to develop a quantitative calibration model which is evaluated for its predictive accuracy using an independent test set, referred to as a validation model (Hutengs et al., 2021; Li et al., 2021; Thomas et al., 2021; Trucano et al., 2006). Predictive modelling that uses a good proportion of soil and plant property data

is known as a supervised technique (Fig. 2.5), whereas predictive modelling using fewer soil and plant property data is known as a semi-supervised technique. Predictive modelling based solely on spectral data is an unsupervised technique (Kuhn and Johnson, 2013).

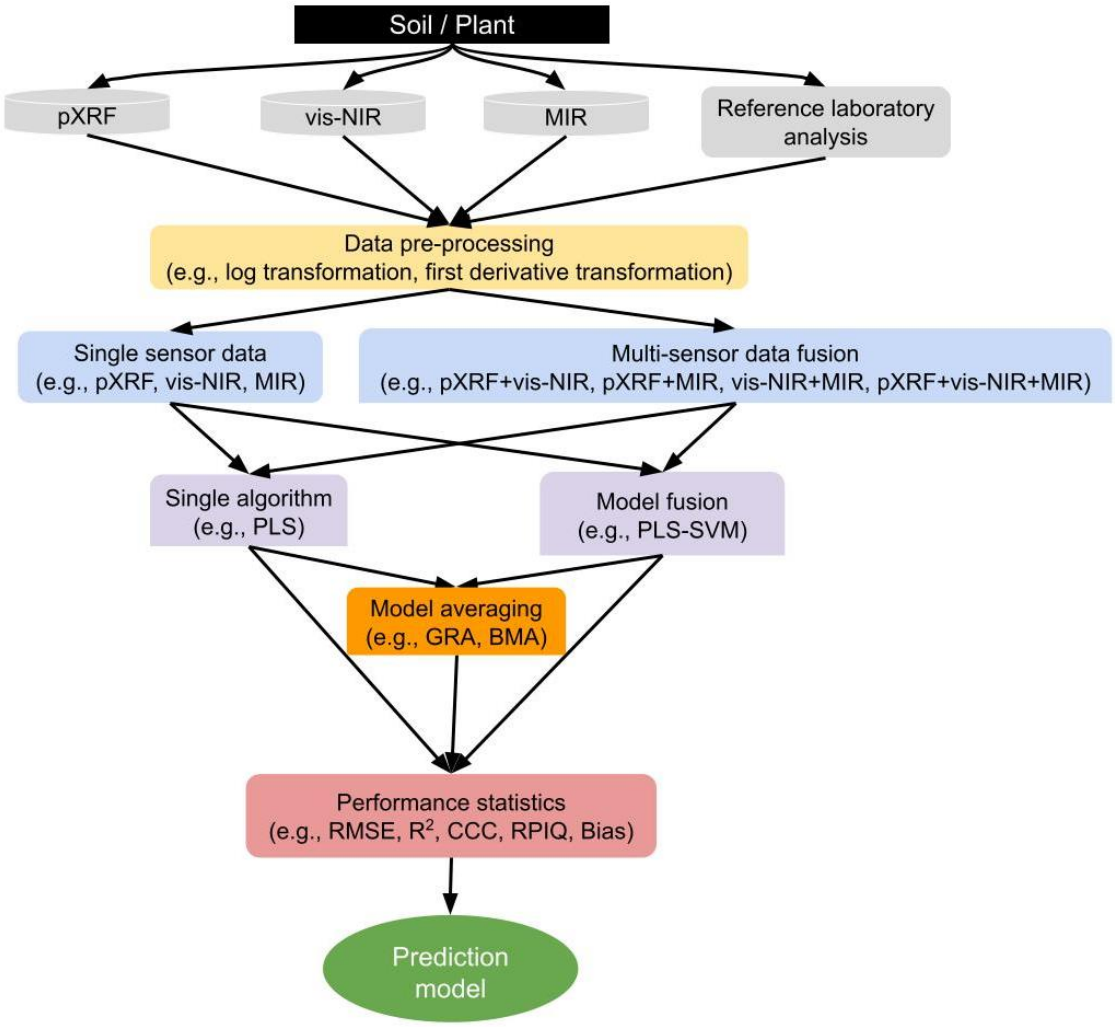


Fig. 2.5 A flow diagram of prediction model development procedure assessing soil and plant properties (including Cd quantification) using proximal sensing techniques including pXRF, vis-NIR, and MIR. '+' sign is for concatenating multi-sensor data.

Table 2.7 Key features of selected proximal sensing techniques used to assess TE concentrations.

Proximal sensing techniques	**portable X-ray fluorescence**	**Visible -near-infrared**	**Mid-infrared**
Wavelength (nm)	0.001−10	350−2500	2500−25000
Principle	• Fluorescence energy is released during the excitation of the element from the ground state. • XRF directly records the signal from TE.	• Records spectral response of overtones and combination of vibrations of molecular bonds and functional groups in the matrix. • Vis-NIR records signals of spectrally active soil components with which TE are associated.	• Records spectral response of fundamental vibrations of molecular bonds and functional groups in the matrix. • MIR records signals of spectrally active soil components with which TE are associated.
Advantages	• Direct TE quantification is possible using built-in calibration model. • Analytical results are comparable with inductively coupled plasma atomic emission spectrometry (ICP-AES) quantification for Mn, Fe, Cu, Zn, Cd and Pb in contaminated soils.	• Many wavelengths in the visible region (400–780 nm) of soil spectra correlate with TE due to their strong bonding with organic matter and iron minerals. • Carbonate (CO_3), AlOH, FeOH, MgOH, S, and hydrated sulphate (SO_4) can be detected.	•Provides information-rich data as organic functional groups have characteristic and well delineated absorption bands in the spectral region. • Calibration data are generic and can be transferred from instrument to instrument.
Limitations	• Can penetrate up to 2 to 5 mm only. • Concentration results may suffer absorption (lower value) and enhancement (higher value) effects due to matrix interference by soil moisture, soil organic matter, and particle size. • Concentration over-reporting due to spectral overlapping. • Recommended for air-dried sieved (2 mm) samples.	• No penetration. • Very low chemical specificity. • Functional groups such as C=O and C-N do not produce a distinct, identifiable spectrum. • Indirect quantification of soil TE based on their relationship with spectrally active soil components (e.g., soil organic matter, Al–, Fe–, and Mn–containing minerals). • Recommended for air-dried sieved (2 mm) samples.	• No penetration. • Available energy in the MIR region decreases substantially with wavelength. • Indirect quantification of TE is based on their relationship with spectrally active soil components (e.g., soil organic matter, Al–, Fe–, and Mn–containing minerals). • Recommended for air-dried sieved (0.5 mm) samples.
Use in TE assessment	• TE can be detected and quantified directly above the limit of detection. • Raw XRF spectra analysis using chemometric methods can quantify lower Cd concentrations than the detection limit of built-in calibration model. • Used for *in-situ* measurement of 4–70 mg Cd/kg soil, 17–3988 mg Cu/kg soil, and 36.3–24187 mg Zn/kg soil. • Not yet used for quantification of TE fractions.	• A low concentration of TE has been predicted by models using vis-NIR spectra based on TE association with spectrally active soil components. • Correlation between soil organic matter and TE has been used to develop TE prediction models. • Used to determine agricultural soil concentration range of 0.01–216 mg Cd/kg soil, 0.2–2397 mg Cu/kg soil, and 0.1–9082 mg Zn/kg soil. • Used for assessing TE fractions.	• Used for predicting the agricultural soil concentration range of 0.005–199 mg Cd/kg soil, 0.2–391 mg Cu/kg soil, and 1.5–4500 mg Zn/kg soil. • Used for carbon fractionation but not yet used for analysing TE fractions.
References	(Caporale et al., 2018; Lemière, 2018; Ravansari and Lemke, 2018; Rouillon and Taylor, 2016; Weindorf et al., 2012; Wu et al., 2012; Xu et al., 2020)	(Chang et al., 2001; Lamine et al., 2019; O'Rourke et al., 2016a; Reeves III and Smith, 2009; Roberts et al., 2018; Shamsoddini et al., 2014; Stafford et al., 2018b; UAE, 2020; Wu et al., 2009; Zhang et al., 2019)	(Dong et al., 2011; Gromski et al., 2015; McCarty et al., 2002; Rajendram and Devey, 2011; Siebielec et al., 2004; Soriano-Disla et al., 2013)

2.4.1 Spectral data pre-processing methods

Spectra pre-processing refers to any mathematical treatment of the raw spectral data before use in the development of a prediction model (Fig. 2.5; Table 2.8). Spectral data pre-processing is required to remove multiple artefacts in the raw spectra including noise (Clingensmith et al., 2019), the effect of light scattering, spectral interference, baseline shift, and to remove irrelevant information (Table 2.8; Angelopoulou et al. (2017)). This step is also useful to correct any differences in soil spectra due to grain size (Shamsoddini et al., 2014). Furthermore, pre-processing enhances wavelength signals relevant to soil properties for predictive modelling (Angelopoulou et al., 2017). Different pre-processing techniques may result in dissimilar predictive accuracy and wavelengths of importance for TE prediction (Liu et al., 2017; Shin et al., 2019). Hence, the choice of spectral data pre-processing has a significant role in TE predictive modelling (Kemper and Sommer, 2002).

Among pre-processing methods, log transformation (Bray et al., 2009) and Savitzky-Golay filters (Savitzky and Golay, 1964) are applied either alone or in combination before other pre-processing (Table 2.8). There are specific pre-processing methods that account for baseline shift (e.g., detrending, derivatives), light scattering effect (standard normal variate, robust normal variate, multiplicative scatter correction, continuum removal, normalisation) or variation from external factors (orthogonal signal correction) and are referred to as first level pre-processing methods (Table 2.8; Campos and Reis (2020)). Among these, first-order derivative transformation is one of the most common pre-processing method to normalise spectra and intensify spectral signals to yield substantially more information on which calibration can be based (Table 2.8; Kemper and Sommer (2002)). In many cases, derivative transformations are applied in combination with Savitzky-Golay filters (Khosravi et al., 2018). Another common pre-processing method is continuum removal, which highlights the energy absorption feature of minerals (Table 2.8; Rathod et al. (2015b)).

Second level pre-processing methods include mean centring and variance scaling (Campos and Reis, 2020). Mean centring (the process of subtracting the mean of each variable from the column of the data matrix) helps to avoid baseline offset in the predictors. Variance scaling, achieved by dividing variables by their respective standard

deviations, is used to impose equal weights on the data sets containing variables with different units and scales (Campos and Reis, 2020). Third level pre-processing method, described by Campos and Reis (2020) as multiblock analysis, can find associations between information from multiple sources and assigns weights to each data block before use in model construction.

There are also pre-processing methods used for specific instruments. These include splice correction for vis-NIR (Stevens and Ramirez Lopez, 2020), constant offset elimination for MIR (Moros et al., 2009), and Coiflet 3 for pXRF spectral data (Li et al., 2017). Before using spectral data as input for algorithms, some other data filtering and separation techniques are used to increase the reliability and robustness of the prediction model.

2.4.2 Outlier detection and sampling methods

For outlier detection in the spectral dataset, Mahalanobis distance is the most used method (Blaschek et al., 2019). Mahalanobis distance accounts for the fact that two spectra may show similarity for properties of two samples due to the highly correlated wavelengths in the spectra (Wadoux et al., 2021a). Other methods including principal component analysis, Euclidian distance, and correlation dissimilarity are also used (Table 2.9).

For separating the sample dataset into training and test sets, different sampling methods are used including the Kennard-Stone sampling, K-means clustering, and conditioned Latin hypercube sampling (Table 2.10). Kennard-Stone sampling is the most common method in chemometric analysis (Table 2.10; Ng et al. (2018)).

Table 2.8 Selected spectral data pre-processing methods for predictive modelling using proximal sensor data.

Artefacts	Pre-processing methods	Description	Aim	Comments	Use in predictive modelling of TE	References
Light scatter (additive or multiplicative perturbations)	Log transformation	Calculates $\log_{10}(1/R)$; where R is the reflectance value.	Remove the scattering effect.	Usually, the first step in vis-NIR spectra pre-processing.	In MIR, used to develop a predictive model based on principal component analysis.	(Bray et al., 2009)
	Continuum removal	Generates new spectral data by dividing the envelope curve of a continuum on raw reflectance spectra.	Effective at isolating specific absorption features, removing the effects of changing slopes and overall reflectance levels.	Removes humidity differences between targets and illumination between shots.	With support vector machine regression model, this pre-processing was chosen for predicting Cd.	(Clark and Roush, 1984; Gholizadeh et al., 2015)
	Robust normal variate	Row-wise scaling operation that removes the spectrum median from all the spectrum variables and divides them by the spectrum robust standard deviation.	Ensures more efficient light scatter corrections.	The percentile approach is more robust than the mean and less sensitive to extreme values.	Not yet applied for TE predictions.	(Guo et al., 1999)
Noise	Savitzky-Golay filters	Local fitting of low-order polynomials and their subsequent use as an efficient filtering scheme.	Reduce the artificial noise caused by the proximal sensor instrument.	An optimal value for Savitzky-Golay filters is determined by trial and error.	The most common pre-processing method applied to predict TE.	(Khosravi et al., 2018; Savitzky and Golay, 1964; Wang et al., 2014)
Baseline shift	Derivatives	Calculates the derivative of the spectra measured as per the variable number (index) or another relevant axis scale (wavelength).	Remove additive constant background effects (1^{st} derivative). Remove baseline linear slope variations and additive effective (2^{nd} derivative).	Because derivatives de-emphasise lower frequencies and prioritise higher frequencies, they tend to accentuate noise.	Increase correlation between reflectivity and TE concentration, limit the influence of partial linearity (1^{st} derivative); Prediction model using 2^{nd} derivative spectra quantified Cd accurately than other pre-processing.	(Brown et al., 2000; Chen et al., 2015; Shamsoddini et al., 2014; Todorova et al., 2014)

Table 2.9 Selected outlier detection methods for predictive modelling using proximal sensor data.

Outlier detection methods	Formulas	Features	References
Mahalanobis distance	$D(x_a,x_b) = \sqrt{(x_b - x_a)C^{-1}(x_b - x_a)^T}$ For two spectra x_a and x_b with b bands, C is the variance-covariance matrix between the spectra.	Accounts for correlation between specific wavelengths in the spectra.	(Ranga Suri et al., 2019; Wadoux et al., 2021a)
Euclidian distance	$D(x_a,x_b) = \sqrt{(x_b - x_a)I^{-1}(x_b - x_a)^T}$ Where I is the identity matrix.	The most common method to find a distance between two points.	
Principal component analysis	$X = TP^T$ Where X is the matrix of size n x b containing the spectra, where n is the number of spectra and b is the number of wavelengths, T = XP is the matrix of scores of size n x d, where d is the number of principal components; and P is the loading matrix of size d x b. The superscript T means the transpose of the matrix.	A statistical technique used to examine the inter-relations among spectra in the dataset and to identify their underlying structure.	
Correlation dissimilarity	Corr diss$(x_a,x_b)=\frac{1-r(x_a,x_b)}{2}$	Correlation dissimilarity is based on Pearson's r correlation coefficient between spectra.	

2.4.3 Predictive modelling

Several algorithms can be used to develop prediction models for quantitative and qualitative assessment of soil and plant properties, including Cd concentration, using regression and classification algorithms (Fig. 2.5; Table 2.11). Some algorithms, such as partial least squares (PLS) regression, are specific for quantitative assessment whereas others such as support vector (SVM), can be used for both quantitative and qualitative evaluation. Algorithms can be divided into linear and non-linear types (Table 2.11; Kuhn and Johnson (2013)). Linear algorithms consider linear relationships between predictors and soil and plant property (e.g., PLS) whereas non-linear algorithms are used when non-linear relationships exist between predictors and target soil and plant property (e.g., neural networks, SVM; Table 2.11). Among non-linear algorithms, tree- and rule-based algorithms use bootstrapping and rules in conjugation, for example, random forest and cubist methods (Kuhn and Johnson, 2013). Furthermore, model fusion approach e.g., PLS-SVM can benefit from the best of two or more algorithms to develop a prediction model (Bao et al., 2017; Blaschek et al., 2019).

Table 2.10 Selected sampling methods to separate training and testing sets for predictive modelling using proximal sensor data.

Sample selection methods	Definitions	Main characteristics	Limitations	References
Random selection	Select samples randomly without considering sample characteristics and distribution.	• The simplest way of selecting samples. • An unbiased method.	More samples are required to achieve the representativeness of the data.	(Ng et al., 2018)
Similarity analysis	Select samples based on spectral similarity by principal component analysis; Define the most similar samples using searching criteria based on the k-nearest neighbours or Mahalanobis distance.	• Project the spectra into low dimensional orthogonal variables. • Compute and evaluate spectral similarity/dissimilarity matrices. • Remove irrelevant spectra from a reference set.	Requires multiple steps.	(Nawar and Mouazen, 2018)
Kennard-Stone	Sequentially select samples with the largest distance in the soil and plant property result data in the training set.	• Sequential and deterministic procedure.	This algorithm picks extreme spectra, producing a good calibration for specific soil properties, but poor calibration for other properties.	(Kennard and Stone, 1969)
Duplex	Like Kennard-Stone, this algorithm selects samples from large multivariate data.	• Allows selecting both calibration and validation samples that are independent. • Can be used when a larger training set is desired than a testing set.	Replicates are divided equally in the training and testing set.	(Snee, 1977)
OptiSim	Optimisable K-Dissimilarity Selection adjusts the balance between representativeness and diversity in the samples.	• Includes maximum and minimum dissimilarity-based selection as special cases. • Mimic representativeness of selections based on hierarchical clustering.	Requires input parameter threshold (ε) which is defined as the minimal distance between "dissimilar" objects.	(Clark, 1997)
K-means clustering	Select samples randomly after stratifying into similar groups (clusters).	• Partitions data into clusters with similar properties, and then random sampling is used to select representative samples from the strata.	This algorithm does not partition the data effectively for a large dataset.	(Ng et al., 2018)
conditioned Latin Hypercube Sampling	Select samples that optimally represent the multivariate distribution of the input dataset.	• Stratify samples in presence of ancillary data. • Provide a more robust sampling algorithm regardless of sampling size.	Availability of complete ancillary information.	(Minasny and McBratney, 2006)

With the availability of multiple proximal sensor instruments in soil laboratories, combining the sensor data from two or more proximal sensor instruments can boost the predictive power achieved from a single sensor (Fig. 2.5; O'Rourke et al. (2016b)). A synergistic combination of two or more proximal sensors can be achieved by: (1) data fusion (Wang et al., 2015), and (2) model averaging (O'Rourke et al., 2016b). The integrated power of multiple sensors either by sensor data fusion or a combination of individual sensor data-based model output can improve accuracy and reduce the uncertainty of prediction models developed for quantitative and qualitative assessment.

2.4.3.1 Single regression algorithms

Partial least squares (PLS) regression is one of the most successfully used algorithms in analysing data from proximal sensor instruments (Nawar et al., 2019). The PLS regression can also be used as a dimensionality reduction technique because it performs matrix decomposition on the spectra (predictors) and soil and plant properties data (target0 to extract latent variables that maximise the covariance between predictors and target properties (Table 2.11). Support vector machine (SVM) is an instance-based algorithm, which typically builds a database using similarity to find the best match and predict matrix properties (Table 2.11; Brownlee (2019)). With SVM, large outliers have a limited effect on the model equation (Kuhn and Johnson, 2013). The inclusion of linear , polynomial, or radial basic function and hyperbolic tangent kernel functions to encompass linear and non-linear functions of the predictors gives this algorithm more predictive power than PLS (Gholizadeh et al., 2015). However, SVM is suitable for limited size datasets, and may result in overfitted models without the possibility of using them for general purposes when non-linear kernels are used and pairwise distances between samples are large (Han and Jiang, 2014). Random forest (RF) is an ensemble learning method that combines Breiman's bagging approach and random selection of wavelengths to construct multiple decision trees (Table 2.11; Breiman (2001)). A random forest model achieves variance reduction by selecting robust and complex decision trees exhibiting low bias. Each tree is selected independently of all previous trees. However, the ensemble nature of random forests makes it difficult to understand the relationship between the predictors and target properties (Thompson, 2019). Random forest may have higher predictive accuracy than PLS but may be of limited use due to the inability to

Table 2.11 Selected algorithms for predictive modelling using proximal sensor data.

Model types	Algorithms	Definitions	Main characteristics	Limitations	Use in TE assessment	References
Linear	Partial least squares (PLS)	A method to relate a single soil and plant property or a matrix of soil and plant properties to predictors.	Maximisation of the covariance between the latent variables and the response.	Requires a high signal-to-noise ratio; Risk of overlooking "real" correlation and sensitivity to the relative scaling of the descriptor variables.	For 0.08–1.44 mg Cd/kg soil, RMSE value was observed below 0.2 mg Cd/kg soil using the PLS model for vis-NIR and MIR.	(Acharjee, 2012; Cramer, 1993; Song et al., 2012)
Non-linear	Support vector machine (SVM)	A pattern recognition method based on statistical learning theory, for small non-linear and high-dimensional samples.	Ability to balance between accuracy attained on a given finite amount of training patterns and generalise to unknown samples.	Choice of the kernel function; speed and size in training and testing.	TE prediction using first-order derivative pre-processing and SVM regression outperformed PLS regression.	(Chen et al., 2015; Cunha et al., 2020; Gholizadeh et al., 2015; Patle and Chouhan, 2013)
Tree- and rule-based	Random forest (RF)	An ensemble learning method that combines a bagging approach and random selection of predictors (wavelengths) to construct multiple decision trees.	Each note is split using the best among a subset of predictors randomly chosen at that node, which is a robust classifier against overfitting.		RMSE=0.08 mg/kg soil for 0.03–6.79 mg Cd/kg soil	(Chen et al., 2015; Cipullo et al., 2019; Minasny et al., 2013; O'Rourke et al., 2016b; Thompson, 2019)
Model fusion	PLS-SVM	PLS latent variables are used as input in SVM regression.	Captures non-linear association of concentrations with soil spectral response		This study	(Blaschek et al., 2019; Hong et al., 2019)
	RF-SVM	RF selected variables important for prediction are used as input in SVM regression.			Not used for TE yet.	(Rustam et al., 2019)

extrapolate outside of the input target soil and plant property data range (Douglas et al., 2019).

2.4.3.2 Model fusion

Model fusion (also known as model hybridisation) refers to a combination of two or more algorithms, used in a sequence while developing a prediction model (Liu et al., 2021; Rustam et al., 2019). Model fusion techniques including PLS–SVM (Bao et al., 2017; Blaschek et al., 2019; Hong et al., 2019), partial model tree (Hao et al., 2016), RF–SVM (Rustam et al., 2019), penalised spline regression–RF (Chakraborty et al., 2015) have all been used for accurate prediction of multiple soil properties (Table 2.11). The PLS-SVM combines features of PLS and SVM to predict soil properties. The PLS performs matrix decomposition of predictor and target property data to extract latent variables (Bao et al., 2017) which contributes in (1) reducing data redundancy (Hong et al., 2019), (2) removing noise (Blaschek et al., 2019), and (3) reducing matrix interference from soil organic matter in XRF spectra. These latent variables can then be used in combination with SVM based regression which considers the non-linear relationship between predictors and target property (Blaschek et al., 2019). However, the potential of model fusion techniques for soil TE prediction remains unexplored (Table 2.11).

2.4.3.3 Data fusion

The spectral response detected by one proximal sensor can be expanded by sensor data fusion of more than one instrument (e.g., a combination of pXRF, vis-NIR, and MIR, or vis-NIR and MIR) to increase predictive accuracy. Data fusion methods can be classified as low-, middle-, and high-level fusion (Table 2.12; Xu et al. (2020)). Low-level data fusion is achieved through the concatenation of spectral data from two or more proximal sensors (Xu et al., 2020). Middle-level fusion involves wavelength selection and dimensionality reduction of an individual or multiple sensor data implementing different techniques including principal component analysis (PCA), genetic algorithm (search-based algorithm; based on the concept of natural selection and genetics), variable importance in projection (VIP; based on the importance score given to each wavelength for prediction of a property in concern) (Table 2.12; Stefansson et al. (2020)). High-level fusion is based on outer product analysis, which is defined as a mathematical combination

of two sensor data generating outer products (Table 2.12; Jaillais et al. (2005)). Low-level fusion can lead to redundancy problems which may reduce overall model performance, whereas high-level fusion is computationally intensive. Therefore, middle-level fusion has become more common for improved predictive modelling (Table 2.12). Data fusion of pXRF and vis-NIR has been used in multiple studies, but examples of the inclusion of MIR with pXRF and vis-NIR are limited (Table 2.12; Li et al. (2021); O'Rourke et al. (2016b)).

Table 2.12 Selected data fusion methods for combining sensor data from different proximal sensing techniques for predictive modelling.

Data fusion methods	Main features	Proximal sensing techniques	TE assessed	Limitations	References
Low-level fusion	Concatenating spectral data of different sensors.	pXRF + vis-NIR	As, Cd, Cr, Ni, Pb	Computationally intensive due to data size; may reduce performance due to the redundant spectra.	(Xu et al., 2020)
		TXRF[1] + MIR	Cu, Zn		(Towett et al., 2015b)
		vis-NIR + MIR	Not yet tested		(Ng et al., 2019)
Middle level fusion	Important spectra contributing to the TE quantification are selected from individual sensors.	pXRF + vis-NIR	As, Cd, Cr, Ni, Pb	Computationally less intensive compared to low-level fusion.	(Xu et al., 2020)
	Principal components from vis-NIR spectra are combined with raw spectra of pXRF.	pXRF + vis-NIR	Pb		(Pozza et al., 2020)
	Concatenating principal components from different sensor spectra.	pXRF + vis-NIR	Not yet tested		(Zhang and Hartemink, 2020)
High level fusion (Outer product analysis)	Emphasise the co-evolution of spectral regions in signals acquired in two spectral regions (sensors).	pXRF + vis-NIR	As, Cd, Cr, Ni, Pb	Computationally intensive due to very large matrix generated; only data from two sensors have been combined.	(Jaillais et al., 2005; Xu et al., 2020)
		vis-NIR + MIR	Not yet tested		(Ng et al., 2019)

TXRF[1] = Total X-ray fluorescence spectroscopy

2.4.3.4 Model averaging

Model averaging is the procedure of combining model outputs based on individual sensors into a single prediction model (O'Rourke et al., 2016b). Model averaging allows the processing of information from two or more proximal sensors, including their associated uncertainty (Abbott, 2014). Model averaging addresses problems such as data redundancy, increasing data size, and longer computation time associated with data fusion (Abbott, 2014). Granger-Ramanathan model averaging (GRA), Bates-Granger or variance weighted averaging (VWA) (O'Rourke et al., 2016b), and Bayesian model averaging (BMA) (Xu et al., 2019a) methods have been used to predict TE concentrations in soils (Table 2.13). In one study, the synergistic use of vis-NIR, MIR, and pXRF through GRA model averaging resulted in the improved predictive accuracy of soil TE concentration when compared with the output of the individual proximal sensors (Table 2.13; O'Rourke et al. (2016a)). Model averaging improves predictive accuracy and robustness (Abbott, 2014). It is achieved by combining multiple models into a single prediction by reducing (1) the dominance of a single model in the final predictions and (2) the likelihood of aberrant prediction (O'Rourke et al., 2016b). However, in all reported studies the level of improvement was limited due to the inclusion of sensors generating indirect methods of prediction relying on outputs from the models developed using a linear or non-linear algorithm (O'Rourke et al., 2016a; O'Rourke et al., 2016b; Pozza et al., 2020; Xu et al., 2020; Xu et al., 2019a).

Granger-Ramanathan model averaging method uses the ordinary least squares combination method to estimate the weights and the intercept for the combination of predictions (Granger and Ramanathan, 1984). The TE concentration can be calculated as multiple linear equations for y after GRA according to Equation 1:

$$y_{GRA} = b + \sum_1^i (W_i + X_i) \quad (1)$$

Where b is the intercept and W_i is the combination weight given to the individual model predictions X_i based on i^{th} sensor (pXRF, vis-NIR, and/or MIR) in the model averaging algorithm.

Table 2.13 Main features and limitations of selected model averaging techniques for predictive modelling using proximal sensing techniques.

Model averaging methods	Main features	Proximal sensing techniques	Use in TE assessment	Limitations	References
Granger Ramanathan model averaging	Ordinary least squares are used to solve the weight of the sensor outcomes.	pXRF, vis-NIR, MIR	For As, RMSE of 3.68 mg/kg compared to 5.25 mg/kg using pXRF.	The level of improvement is limited by the algorithm used to develop the prediction model.	(O'Rourke et al., 2016a)
		pXRF, vis-NIR	For Cr, RMSE of 10.33 mg/kg compared to 12.19 mg/kg using pXRF.		(O'Rourke et al., 2016b) (Pozza et al., 2020; Xu et al., 2020)
Bayesian model averaging	Considers the uncertainty of the model itself.	pXRF, vis-NIR, MIR	Similar performance as GRA.	Compared to GRA, complexity in the algorithm.	This study
Variance weighted averaging	Assumes that the variances from each model outcome are uncorrelated.	pXRF, vis-NIR	For cobalt, RMSE of 2.05 mg/kg compared to 2.91 using vis-NIR.	Underperformed compared to GRA	(O'Rourke et al., 2016b)

The Bayesian model averaging (BMA), a linear model, posterior distribution (p) of the final TE concentration (y_{BMA}) can be expressed as Equation 2 (Clyde, 1999):

$$p(y_{BMA}|y_{measured}) = \sum_1^i p\,(y_{BMA}|y_{measured}, y_i)p(y_i|y_{measured}) \quad (2)$$

Where $y_{measured}$ is the measured TE concentration, and y_i is the TE concentration output given by the prediction model based on the ith proximal sensor (pXRF, vis-NIR and/or MIR) for the $y_{measured}$ concentration.

BMA posterior distribution of y_{BMA} is a weighted average of the posterior distribution of y_{BMA} under each of the sensors, weighted by their posterior model probabilities. The posterior model probability of y_i is expressed as Equation 3 (Clyde, 2020):

$$p(y_{BMA}|y_{measured}) = \frac{p(y_{measured}|y_i)p(y_i)}{\sum_{i=1}^n p(y_{measured}|y_i)p(y_i)} \quad (3)$$

where $p(y_{measured}|y_i)$ is the integrated likelihood of y_i which can be calculated by *JZS* prior distribution for regression coefficients (Clyde, 2020).

2.4.3.5 Memory based learning algorithms

Memory based learning (MBL) algorithms use single regression algorithms to develop prediction model specific to each sample (Shenk et al., 1997). These algorithms including locally weighted regression (Naes et al., 2002), LOCAL (Shenk et al., 1997), SBL (Ramirez-Lopez et al., 2013), and RS-LOCAL (Lobsey et al., 2017), are implemented to quantify local sample soil properties using a spectral library (Ng et al., 2022b). The method of selection of SSL subset to predict the property of a new sample may differ as per MBL algorithm. To select calibration samples, the LOCAL algorithm uses Mahalanobis distance, whereas in SBL distance matrices calculated in the principal component space is used (Ramirez-Lopez et al., 2013; Shenk et al., 1997).

2.4.4 Identifying spectral regions of importance for Cd prediction

Both PLS loadings and variable importance in projection (VIP) scores given to wavelengths in the optimal prediction model show relevance of the specific spectral regions to assess the association of target soil or plant property (e.g., Cd concentration) with matrix components (e.g., soil organic matter or plant cell structure). PLS loadings define the linear combination of the predictors that maximises covariance with the target property (Kuhn and Johnson, 2013). To determine PLS loadings, first PLS determines underlying (latent) relations among predictors (sensor data), which are highly correlated to the target property (Kuhn and Johnson, 2013). This relationship is numerically summarised as a vector of weights also known as a "direction". Then, predictor data are orthogonally projected onto the "direction" to generate scores. The scores are then used to generate loadings, which measure the correlation of the score vector to the original predictors.

PLS VIP scores summarise the influence of the individual wavelength on the PLS model. The VIP scores are calculated as the weighted sum of squares of the PLS weights, which consider the amount of explained target property (e.g., Cd concentration) variance in each

extracted latent variable (Farrés et al., 2015). The scores can be scaled to have a maximum value of 100 for ease of understanding (Kuhn et al., 2021).

2.4.5 Performance statistics assessing quantitative predictive models

The predictive accuracy of constructed models can be assessed and compared using a combination of several performance statistics: root mean square error (RMSE, Equation (4)), normalised RMSE (nRMSE, Pullanagari et al. (2016); Equation (5)), coefficient of determination (R^2, Equation (6)), ratio of performance to interquartile distance (RPIQ; Bellon-Maurel et al. (2010); Equation (7)), Lin's concordance correlation coefficient (CCC; Lin (1989); Equation (8)), and bias (Equation (9)).

$$\text{RMSE} = \text{predicted} - \text{measured} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} \quad (4)$$

$$nRMSE = \frac{RMSE}{Mean} \quad (5)$$

$$R^2 = 1 - \frac{\text{Sum of square}_{\text{residual}}}{\text{Sum of square}_{\text{total}}} = 1 - \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{(y_i - \mu_y)^2} \quad (6)$$

$$\text{RPIQ} = \frac{Q_3 - Q_1}{\text{RMSE}} \quad (7)$$

$$\text{CCC} = \frac{2rs_x s_y}{s_x^2 + s_y^2 + (\mu_x - \mu_y)^2} \quad (8)$$

$$\text{Bias} = \frac{\sum_{i=1}^{n}(y_i - x_i)}{n} \quad (9)$$

For a dataset of n samples, x_i is a measured TE concentration of the i^{th} sample with an associated predicted value y_i, μ_x is the mean value for measured concentrations and μ_y is the mean value for predicted concentrations; and s_x^2 and s_y^2 are the corresponding variances. In Equation (7), Q_1 and Q_3 are the first and third quartile values of measured TE concentrations. In Equation 8, r is the correlation coefficient between the measured and predicted TE concentrations.

Among statistical parameters, RMSE shows overall accuracy of the prediction model whereas Bias measures systematic error in the predictions made by the model. Both

measures depend on the scale of the data (Wadoux et al., 2021a). Whereas nRMSE is independent of the scale of the data, so easier to understand and compare accuracy of prediction models based on different scale dataset (Pullanagari et al., 2016). The CCC measures agreement between measured and predicted values in the prediction model whereas R^2 value measures the percentage of target property (measured values) variance explained by the prediction model (Lin, 1989). The RPIQ can be used to compare performance of difference models to evaluate improvement (Bellon-Maurel et al., 2010).

2.5 Using large spectral libraries for local soil Cd quantification

A soil spectral libraries (SSL) is a collection of spectral, physical, chemical, biological, and spatial information on soil samples with the record of the relationship between soil properties and their spectral response (Viscarra Rossel et al., 2016). Such SSLs are the basis for the wider application of these proximal sensing techniques to accurately estimate soil properties (Shepherd and Walsh, 2002). Availability of SSLs at large scales e.g., regional/national, continental, and global, can potentially reduce the need to collect large numbers of soil samples and to analyse these using reference laboratory methods to develop calibration models for predicting soil properties locally (Table 2.14; Brown (2007); Sila et al. (2016); Viscarra Rossel et al. (2016)). In this way, a robust SSL can reduce or completely mitigate the need for routine wet chemistry analysis as a part of model development (Rossel et al., 2008; Shepherd and Walsh, 2002). The robustness of large SSLs is directly related to the data size, representative pedo-diversity, and coverage of soil properties (e.g., the concentration range of TE as Cd) of interest (Viscarra Rossel et al., 2016). However, large SSLs may require additional steps such as (1) selection of a large SSL subset, (2) spiking with local samples, (3) giving weights to local samples, and/or (4) selection of algorithm before model development and predict the property of interest in local samples (Brown et al., 2006; Greenberg et al., 2022; Ng et al., 2022a; Wetterlind and Stenberg, 2010).

The selection of a suitable subset of samples from the large SSL can be made based on (1) common characteristics e.g., similarity of land use (Moura-Bueno et al., 2020), and/or (2) spectral similarity e.g., using principal component analysis (PCA) to find samples from the large library in the closer spectral distance with the targeted local samples

(Nawar and Mouazen, 2018). Li et al. (2020) found that reducing the geographical coverage of a large-scale SSL increases the accuracy of a prediction model (Table 2.14). Spiking refers to adding representative samples from a local area into the SSL (e.g., regional, national, continental, or global SSLs) or its subset to develop a prediction model (Brown, 2007). Two methods have been proposed to add local samples to the SSL or its subset: (1) spectral similarity, as assessed by PCA (Nawar and Mouazen, 2017b); and (2) the use of concentration values (Table 2.14; Guerrero et al. (2010)). When using concentration values, methods as conditioned Latin hypercube sampling, Kennard-Stone sampling, Mahalanobis distance (distance between a sample and a distribution), and Shannon H index (measure diversity in the dataset) methods have been considered to make the sampling effort unbiased (Table 2.10; Guerrero et al. (2010); Li et al. (2020)).

Guidelines have been proposed for considering the appropriate number of samples that should be added to large SSLs to improve predictions such as Li et al., (2020) stated that 15% should be added to district level SSL, while Nawar and Mouazen (2017) suggested 70% for continental level SSL. Another method to improve the performance of a prediction model developed using a SSL to quantify local samples is giving extra weight to the spiked local samples (Greenberg et al., 2022). Sankey et al. (2008) showed better large-scale model performance by giving higher weight to local samples when compared to a model that contains no local samples.

Visible-NIR and MIR proximal sensor-based SSLs have been used to predict soil properties as soil organic carbon and total nitrogen locally (Table 2.14; Li et al. (2020); Sanderman et al. (2021)). For example, spiking has been applied to quantify soil carbon content (Kuang and Mouazen, 2013). Testing how a pXRF sensor-based large SSL (e.g., regional) improve the prediction of a soil property as Cd concentration locally would need further research. Moreover, the large SSL, compiling data from different sensors (pXRF, vis-NIR, and MIR) could potentially be used to assess multiple soil properties including Cd concentration at a local scale.

Different algorithms have been tested to predict local soil properties using a large SSL (Table 2.14). The prediction model developed using a SSL or its subset spiked with a representative local samples as input for PLS regression has been shown to generate optimal model to quantify local soil properties (Table 2.14; Brown (2007); Nawar and Mouazen (2017b); Wetterlind and Stenberg (2010)). Among different algorithms,

LOCAL algorithm could potentially be applied to assess low Cd concentrations using a large SSL.

Table 2.14 Selected references quantifying local soil properties using soil spectral libraries (SSLs).

Soil properties	SSLs (size)	Local samples (size)	Algorithms	RMSE (g/kg)	R^2	References
vis-NIR						
Soil organic carbon	NRCS[1], US global spectra database (4184)	Uganda catchment area (1845)	Boosted Regression Trees (BRT)	2.6	0.55	(Brown, 2007)
		Montana, USA (225,52,54)	PLS	3.8, 6.7, 26.2	0.89	(Sankey et al., 2008)
	European continental dataset (529)	Hessleskew (122); Hagg (149)	Multivariate Adoptive Regression Splines (MARS)	0.6	0.98	(Nawar and Mouazen, 2017b)
	European Continental dataset (LUCAS[2]) (19969)	Part of LUCAS as a local library	PLS	3.6	0.84	(Nocita et al., 2014)
	Sweden (396)	Four farms (81, 94, 103, 112)	PLS	1.9–4.4	0.72–0.87	(Wetterlind and Stenberg, 2010)
	China Zhejiang Province (714)	Shoxing city (60), Jinhua city (66)	PLS	1.88–2.10	0.86–0.90	(Li et al., 2020)
	France (2126)	Le Peyne catchment (144)	fast Fourier transform local weighted (FFT-LW)	3.2	0.5	(Gogé et al., 2014)
	Global SSL (17928)	Australian farm (599), New Zealand farm (285)	PLS	0.48 (Australian farm); 1.16 (New Zealand farm)	0.78–0.84	(Kuang and Mouazen, 2013)
Total nitrogen	Spain (1040)	4 sites (44,91,80,45)	PLS	<0.80	>0.90	(Guerrero et al., 2010)
MIR						
Soil carbon fractions	US National dataset (14000)	Australian dataset (565)	Local PLS	1.37	0.89	(Sanderman et al., 2021)
Total carbon	Africa Soil information services (1902)	Central Africa (80)	Memory based learning (MBL)	0.02	0.79	(Summerauer et al., 2021)

[1]NRCS=National Resources Conservation Service
[2]LUCAS=Land Use and Coverage Area frame Survey

2.6 Summary and research gaps

The use of proximal sensing techniques to determine total soil Cd concentrations in New Zealand agricultural soils has the potential to improve the scale and scope of long-term regular monitoring of soil Cd required under the framework of TFMS (Section 2.1.4). To date, proximal soil sensing techniques have focused on the potential to quantify high total soil Cd in the contaminated lands and mining areas (Nawar et al., 2019), which are relatively low in agricultural soils (<3 mg Cd/kg soil; Tables 2.1 and 2.2). The direct quantification of agricultural total soil Cd using pXRF alone is hampered by the low detection limit for Cd associated with this technique (>3 mg Cd/kg soil; Table 2.6; Lemière (2018)). However, the raw XRF spectra can be used to develop calibration models for local samples that have lower limits of detection (O'Rourke et al., 2016a). Site specific soil properties must be considered in model development as soil moisture, soil organic matter, and particle size can all interfere with and attenuate X-ray signals limiting the accuracy of predictions (Section 2.3.1; Ravansari et al. (2020)).

Inexpensive TE monitoring using proximal sensors requires development of a robust SSL including enough samples representing soil orders, parent materials, climatic conditions, and land uses from across the country (Section 2.5). Geochemical soil baseline surveys conducted at a regular spacing across the country can be the representative source of soil samples for constructing a SSL. In New Zealand, GNS Science, a crown research institute conducting research in geological and earth system processes, conducts such geochemical soil baseline surveys collecting samples at regular 8 km spacing and reference laboratory analyses are performed to quantify the total soil concentration of 65 elements, including Cd (Martin et al., 2016). These soil samples are available for scanning by proximal sensors to develop a national scale SSL. The SSLs have been developed for global scale (using vis-NIR) and national scale (MIR) but there is no report of a SSL based on XRF spectra (Gogé et al., 2014; Viscarra Rossel et al., 2016). There is a good opportunity for SSLs to include information from three proximal sensors: pXRF, vis-NIR, and MIR to contribute to reliable and cost-effective monitoring of TE at low concentrations.

A robust SSL developed from regional scale surveys could improve the future prediction of Cd concentration across areas characterised by soil diversity and contrasting Cd concentrations characteristic of farm-scale mapping. Calibration models containing large

scale SSLs have been used to quantify soil properties, including soil carbon, at a local scale (Nocita et al., 2014), and these can be further developed for Cd quantification at a farm scale. With farm-scale Cd concentration quantified, a site-specific Cd management strategy can be implemented effectively. This would be of significant benefit to the ongoing performance of the TFMS to manage Cd levels in New Zealand agricultural soils.

In comparison, reflectance spectra (vis-NIR, MIR) will only generate indirect information about TE concentration as these techniques detect the association of TE with spectrally active soil components (e.g., soil organic matter, Al–, Fe–, and Mn–containing minerals) (Sections 2.3.2 and 2.3.3; Wu et al. (2010)). Studies on the predictive performance of individual proximal sensors (pXRF, vis-NIR and/or MIR sensor) have mainly focused on contaminated soils with elevated TE concentrations (Caporale et al., 2018; Kemper and Sommer, 2002; Reeves III and Smith, 2009) and TE prediction accuracy can be improved by implementing chemometric approaches including model fusion techniques, such as PLS-SVM to reduce noise, outlier influence, and high dimensionality overfitting effect, and capture non-linear relation between TE concentrations and spectra (Section 2.4.3.2; Bao et al. (2017); Hong et al. (2019); Blaschek et al. (2019)). The pXRF can be combined with reflectance spectroscopy techniques (vis-NIR and/or MIR) either by data fusion (Section 2.4.3.3) or model averaging (Section 2.4.3.4) for improved prediction (Li et al., 2021; O'Rourke et al., 2016b; Xu et al., 2020).

The use of proximal sensors to assess total soil Cd, Cd fractions (especially Cd present in fractions available for plant uptake) and plant Cd concentrations can directly contribute to the aim of TFMS to reduce further Cd accumulation in New Zealand agricultural soils and the subsequent risks to the food chain through plant Cd uptake (CWG, 2011). The contribution of both total soil Cd and Cd fractions to plant uptake varies as a function of edaphic and environmental factors, and plant characteristics (Sections 2.1.2 and 2.1.5; Gray et al. (1999b); Stafford et al. (2018c)). Studies have been conducted to quantify soil TE fractions using vis-NIR sensors (Chakraborty et al., 2017; Cipullo et al., 2019), whereas MIR or pXRF sensor could also potentially ease assessing TE fractions. In the case of measuring plant Cd, pXRF or vis-NIR has been used to quantify plant TE (e.g., Cd, Fe, Mn, Pb, Zn) at toxic concentrations and that induce stress in plants (Feng et al., 2019; McGladdery et al., 2018). There is a research opportunity to investigate whether proximal sensing could predict Cd concentration in plant samples (e.g., chicory roots and aboveground biomass; chicory is an important forage herb in New Zealand (Stafford et

al., 2016)), as well as Cd accumulation at non-toxic soil Cd concentrations. Controlled experiments which collect spectral data from soils with increasing Cd concentration (e.g., above concentrations detected in New Zealand; Table 2.1) and with contrasting soil properties covering the expected pedo-diversity within a target location, can contribute to the development of a more representative spectral library which can be used to quantify plant Cd using proximal sensing techniques (Rathod et al., 2013; Sridhar et al., 2007; Viscarra Rossel et al., 2016).

Chapter 3

Quantification of multiple soil trace elements by combining portable X-ray fluorescence and reflectance spectroscopy

A manuscript from this Chapter has been published in the journal Geoderma:

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

Abstract

Several combinations of proximal sensors, pXRF, vis-NIR, and MIR spectroscopy were assessed to quantify multiple soil TE with reliable accuracy. A total of 622 topsoil samples (0–20 cm depth) collected at a regular 8 km spacing across southern New Zealand were analysed for TE concentrations by a reference laboratory method and scanned using pXRF, vis-NIR, and MIR. The laboratory results and spectral information were used to develop robust models which enabled the prediction of soil concentrations of arsenic (As), cadmium (Cd), chromium (Cr), copper (Cu), mercury (Hg), nickel (Ni), lead (Pb), and zinc (Zn). Chemometric approaches including spectral data fusion, PLS-SVM model fusion, and model averaging were tested to obtain optimal quantitative TE predictions in the samples. Validation of PLS-SVM based models on held-out samples showed the combination of pXRF and MIR performed optimally for predicting As, Cr, and Pb concentrations with CCC of 0.97 and RMSE of 0.57 mg As/kg, 1.14 mg Cr/kg, and 1.05 mg Pb/kg, respectively. Fusion of pXRF and vis-NIR data performed optimally for quantifying Cu, Ni, and Zn concentrations with CCC of 0.98 and RMSE of 1.09 mg

Cu/kg, 0.63 mg Ni/kg, and 3.90 mg Zn/kg, respectively. For Cd and Hg, Granger-Ramanathan model averaging of outputs from PLS-SVM models using pXRF, vis-NIR, and MIR data performed optimally with CCC>0.80 and RMSE of 0.03 mg Cd/kg and 0.01 mg Hg/kg, respectively. This study showed that quantitative prediction models for As, Cd, Cr, Cu, Hg, Ni, Pb, and Zn could be successfully implemented at local- and national-scale and for long-term monitoring of soil TE at concentrations below pXRF detection limits and with reduced matrix interference from organic matter than from individual techniques alone.

3.1 Introduction

Understanding TE abundance and variation in natural ecosystems is important for environmental management, agricultural production, and for maintaining human health (Keesstra et al., 2016; Shrestha et al., 2021; Stojsavljevic et al., 2021). Arsenic (As), cadmium (Cd), chromium (Cr), copper (Cu), mercury (Hg), nickel (Ni), lead (Pb), and zinc (Zn) occur naturally in soils at concentrations that typically cause no environmental or health concerns (Hooda, 2010). Some TE, such as Cu and Zn, are essential elements for living organisms, and deficiency in these elements can cause health problems (Marschner, 2012). Background concentrations of these TE in soil have increased in many places worldwide (both urban and rural) as a function of geogenic and anthropogenic activities (Martin et al., 2016; Nriagu and Pacyna, 1988; Turnbull et al., 2019). Common geogenic pathways for increasing TE input into soils are geothermal activity, volcanic eruptions, weathering, and soil erosion (Kabata-Pendias, 2010), while anthropogenic activities include phosphate fertiliser and pesticide application to agricultural land, waste disposal, vehicle emission, and mining and refining activity (Harvey et al., 2017; Morgan, 2010; Tay et al., 2021). Elevated TE concentrations in soil may enhance TE bioavailability, leading to their accumulation in plants and transfer through the food chain, potentially causing environmental, economic, and health risks (Alloway, 2013).

Geochemical soil baseline surveys have been used worldwide to assess and quantify soil TE contamination at regional- to continental scales (de Caritat and Cooper, 2016; Li et al., 2014; Martin et al., 2016; Reimann and de Caritat, 2012; Salminen et al., 2005; Turnbull et al., 2019). Soil TE quantification traditionally uses wet chemistry analytical techniques (e.g., inductively coupled plasma mass spectrometry), which are expensive

and time demanding (Nocita et al., 2015). Given the importance of TE monitoring such as Cd monitoring under the TFMS implemented as a part of the cadmium management strategy (CMS) in New Zealand (CWG, 2011), the use of proximal sensors has been proposed as an alternative soil TE testing procedure (Shepherd and Walsh, 2002), and success has been reported for measuring As, Cd, Cr, Pb, and Zn in contaminated soils (Cheng et al., 2019; Niazi et al., 2015; Xu et al., 2019b).

The current study investigated three proximal sensing techniques, pXRF, vis-NIR, and MIR. The pXRF measures TE concentrations directly, while vis-NIR and MIR quantify TE based on their relationship with spectrally active soil components (e.g., soil organic matter and Al–, Fe–, and Mn–containing minerals) (Lemière, 2018; Soriano-Disla et al., 2014; Wu et al., 2010). The effectiveness of pXRF as an analytical tool is based on (1) the built-in calibration model developed using spectral response of the limited concentrations range and standard matrices or (2) analysis of the raw spectra with chemometric methods (O'Rourke et al., 2016a) and the application of pXRF to quantify TE in the soil is constrained by (1) detection limits rarely approaching those of laboratory-based techniques and (2) soil matrix interferences due to soil moisture content, soil organic matter, and particle size distribution (Padilla et al., 2019; Ravansari and Lemke, 2018; Rouillon and Taylor, 2016). Therefore, there are challenges associated with the use of pXRF to estimate multiple TE at low (background) concentrations. Interferences can be overcome by combining pXRF spectra with vis-NIR and/or MIR spectra using chemometric methods (data fusion or model averaging). These approaches are traditionally employed one technique at a time and/or with a focus on an individual TE (e.g., Cu or Pb) (O'Rourke et al., 2016b; Pozza et al., 2020).

The simple concatenation of raw spectra from different sensors causes data redundancy, limiting improvements in predictive accuracy (Liu et al., 2021; O'Rourke et al., 2016a; Soriano-Disla et al., 2014). Wavelength selection, dimensionality reduction, and spectra exploration methods (e.g., principal component analysis, outer product analysis) have been used to combine multiple sensor data (Benedet et al., 2020; Javadi et al., 2021) for improved TE prediction in contaminated sites (Gholizadeh et al., 2021; Xu et al., 2020). Alternatively, PLS performs matrix decomposition on the spectral and concentration data to extract latent variables (reducing noise) (Blaschek et al., 2019). These latent variables can then be used in combination with SVM based non-linear regression, which has been successfully implemented to quantify soil properties such as organic matter content and

available water-holding capacity (Bao et al., 2017; Blaschek et al., 2019; Hong et al., 2019). However, this PLS-SVM model fusion technique is yet to be tested for estimating background soil TE concentrations. Similarly, multiple sensor data can be combined by applying model averaging methods such as GRA to quantify TE (O'Rourke et al., 2016b; Pozza et al., 2020). Xu et al. (2019a) used the BMA technique to assess the organic matter and total nitrogen in soils, but this approach remains untested to predict TE.

In the present study, a set of soil samples from a geochemical baseline survey covering southern New Zealand were used to test different approaches to determine soil TE concentrations. Spectral response information from pXRF, vis-NIR, and MIR applied in different combinations and with advanced chemometric methods (data fusion, model fusion, and model averaging) were assessed to: 1) quantify multiple soil TE (As, Cd, Cr, Cu, Hg, Ni, Pb, and Zn) using a systematic combination of information from proximal sensors (pXRF, vis-NIR, and MIR) via data fusion or model averaging, and PLS-SVM model fusion, 2) benchmark TE concentrations against the built-in pXRF calibration models, and 3) establish recommendations for the combination of spectra and chemometric methods estimating soil TE concentrations.

3.2 Materials and methods

3.2.1 Study area, soil sampling, and chemical analyses

Soil samples (n=622), collected by GNS Science as a part of the baseline geochemical survey, were obtained from 0 to 20 cm depth from sites spaced 8 km apart covering c. 40,000 km^2 between sea level and 2000 metres above sea level within the Otago and Southland provinces of southern New Zealand (Fig. 3.1; Rattenbury et al. (2018)). Soil samples were taken from the centre and corners of a 20 m square and bulked in the field (Martin et al., 2016). The study area includes diverse geological parent materials, climatic conditions, land uses, and soil TE concentrations (Martin et al., 2016; Rattenbury et al., 2014). The soil sample sites included 47% of arable land (Appendix 3.1). Dominant soil types include Brown (53%) and Pallic soils (24%) (following the New Zealand Soil Classification of Hewitt et al. (2021), Appendix 3.1). Reference laboratory analysis for a suite of elements was conducted on a subsample (0.5 g) that was digested with 10 ml of a solution containing equal parts of concentrated HCl, HNO_3, and H_2O (1:1:1) for an hour

in a heating block or hot water bath (Martin et al., 2016). The digested aliquot was analysed in a PerkinElmer ELAN 9000 inductively coupled plasma mass spectrometer (ICP-MS) to quantify total concentrations of Al, As, Cd, Cr, Cu, Fe, Hg, Ni, Pb, and Zn (Gazley et al., 2020; Martin et al., 2016). Soil total carbon (C) concentrations were measured using a sub-180 µm subsample (5 g) in a LECO CS230 element analyser (Martin et al., 2016).

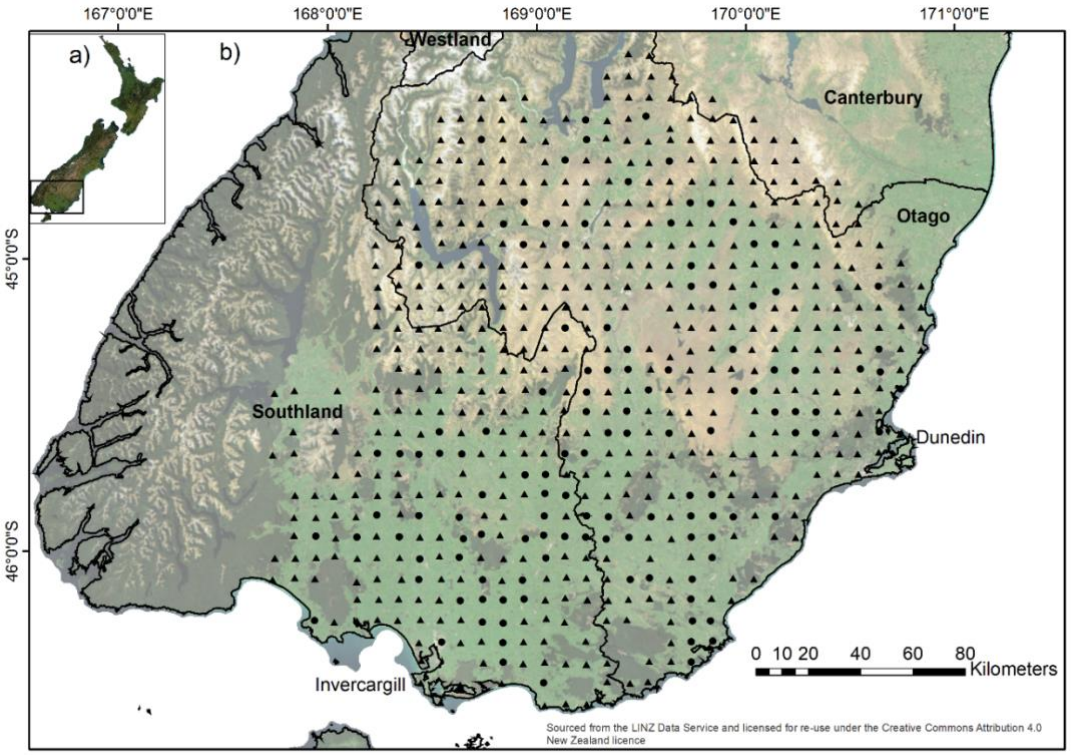


Fig. 3.1 A location diagram showing a) the sampling regional area within New Zealand and b) the detailed site locations (triangles and dots) spaced approximately 8 km apart in southern New Zealand. The site locations are separated into a training set (black triangles) and a testing set (black dots) using the Kennard-Stone sampling method. Provinces are indicated in bold text and cities are in regular font. (Map source: LINZ Data Service, 2021).

3.2.2 Sample preparation

Air-dried sub-2 mm representative subsamples were scanned using three proximal sensing techniques: pXRF, vis-NIR, and MIR (Section 2.3). For pXRF, a cylindrical plastic cup 22 mm height by 14 mm in diameter was filled with air-dried, sub-2 mm soil and covered with a 4 µm thick polypropylene transparent film (Fig. 2.4a-f). For vis-NIR, soil samples were scanned in triplicate. For MIR, soil samples were milled using a Retsch mortar grinder (RM200, Germany) for 30 seconds. Homogenised, finely-ground samples (four replicates per sample) were placed in an aluminium plate with 48 wells and vertical gutters in-between lines of 8 wells (Fig. 2.4m).

3.2.3 Spectral data collection and pre-processing

An Olympus Vanta C series pXRF spectrometer containing a rhodium anode was used to measure XRF spectra (0–40 keV) in *Geochem* mode (*2 beams*) (Fig. 2.4e). This instrument reaches 40 keV while in operation for *beam 1* and 9.98 keV for *beam 2*. *Beam 1* is set for quantifying "heavy elements" including As, Cd, Cr, Cu, Fe, Hg, Ni, Pb, and Zn concentrations. *Beam 2* is to quantify "light elements" including Al concentrations. The silicon drift detector has the expected resolution of 133 eV, recording fluorescence response for 2048 wavelengths. A stainless-steel alloy chip was used to check the built-in calibration of the pXRF instrument. To assess the concentration measurement performance of the built-in pXRF calibration, a standard reference material (SRM) SRM2711a of the National Institute of Standards and Technology was scanned every 20 samples (Fig. 2.4a). The average recovery was calculated using the equation (pXRF measurement/SRM reference value [recovery]) for each TE (Weindorf and Chakraborty, 2020). Recovery of TE concentrations in SRM2711a were As [0.78], Cd [1.00], Cr [2.25], Cu [1.09], Hg [1.73], Ni [1.39], Pb [0.97], and Zn [1.02]. Built-in calibration for the pXRF instrument has detection limits of 2 mg As/kg, 5 mg Cd/kg, 15 mg Cr/kg, 5 mg Cu/kg, 1 mg Hg/kg, 5 mg Ni/kg, 3 mg Pb/kg, and 3 mg Zn/kg soil (Table 2.6). Exposure time for *beam 1* and *beam 2* was 60 seconds each, with measurements taken once per sample. For this study, *beam 1* raw spectra were used after removing the first 24 wavelengths data with zero spectral response values. The TE concentrations reported by the built-in pXRF calibration model were also exported for comparison.

An ASD FieldSpec 3 spectroradiometer (Analytical Spectral Devices Inc., Boulder, Colorado, USA) fitted with a contact probe containing a 4.5 W halogen bulb as a light source was used to record the vis-NIR spectral response of the prepared soil samples between 350 and 2500 nm (Fig. 2.4j). The instrument was calibrated using a Spectralon white reference panel after every ten spectra measurements, using a procedure described by Blaschek et al. (2019) (Fig. 2.4g). The reflectance spectra recorded for each scan were an average of 50 internal readings within the scanning period. Reflectance spectra were then imported into the R statistical environment (R Core Team, 2021). Pre-processing of the spectra included splice correction to remove the artefactual steps occurring at the interface between the three detectors covering the instrument spectral range. The noisiest part of the spectra (350 to 399 nm) was also discarded. The resulting reflectance spectra were then transformed to pseudo-absorbance using the $\log_{10}(1/R)$ formula, where R is the reflectance, then derived and smoothed using a Savitzky-Golay filter (first-order derivative, window size of nine and second-order polynomial) (Savitzky and Golay, 1964).

A Fourier-transform infrared spectrometer (Vertex 70, Bruker, Germany) equipped with a microplate reader extension for high throughput screening infrared spectroscopy equipment (HTS-XT, Tensor II, Bruker, Germany) was used to record MIR spectral response (Fig. 2.4n). The spectral range of the instrument covers 7498–600 cm^{-1} (i.e., 1333–16666 nm). Liquid nitrogen was refreshed at the start of each session and the instrument was tested for performance qualification via the OPUS validation programme using a validation plate (Fig. 2.4l). Quality control was performed by scanning a permanent standard gold sample embedded in the microtiter plate (Fig. 2.4m) along with three organic-rich mineral reference soil samples on each plate. Raw MIR spectra (3578 wavelengths) were exported using OPUS Lab version 7 software (Bruker, Germany), then derived and smoothed using a Savitzky-Golay filter (first-order derivative, window size of nine and second-order polynomial) (Savitzky and Golay, 1964).

The Mahalanobis distance was used to remove outliers from the concatenated spectral dataset. Subsequently, the Kennard-Stone sampling method (Table 2.9) was used to separate the dataset into a training set (80%) and a testing set (20%), independently for each TE (Kennard and Stone, 1969). These same training and testing sets were maintained for prediction model development using three proximal sensors, data fusion, model

fusion, and model averaging. These steps were performed using the R package *prospectr* (Stevens and Ramirez Lopez, 2020).

3.2.4 Modelling framework

The R statistical environment (R Core Team, 2021) was used to perform all chemometric analysis. The spectral data and reference laboratory results were used to develop calibration models based on training sets. Then, the models were validated by using the hold-out testing sets. All concentration values were log-transformed then mean centred and variance scaled before use in the predictive modelling to avoid detrimental effects of the skewness observed in the TE data (Section 2.4.1). During model development, hyper-parametrisation of the different prediction methods tested was carried out using 10-fold repeated cross-validation (18 repeats), with the optimal set of parameters chosen based on the empirically optimal model yielding the smallest RMSE value.

Prediction models were developed using either PLS or a model fusion of PLS with support vector machine (SVM) (Section 2.4.3.2). The PLS regression used the R package *pls* (Mevik et al., 2020), wrapped in the *caret* package (Kuhn et al., 2021). For PLS-SVM model fusion, the optimal set of latent variables extracted from a PLS was then used as inputs for a SVM model (Blaschek et al., 2019) using a polynomial function as implemented in the R package *kernlab* (Karatzoglou et al., 2004) (Fig. 3.2).

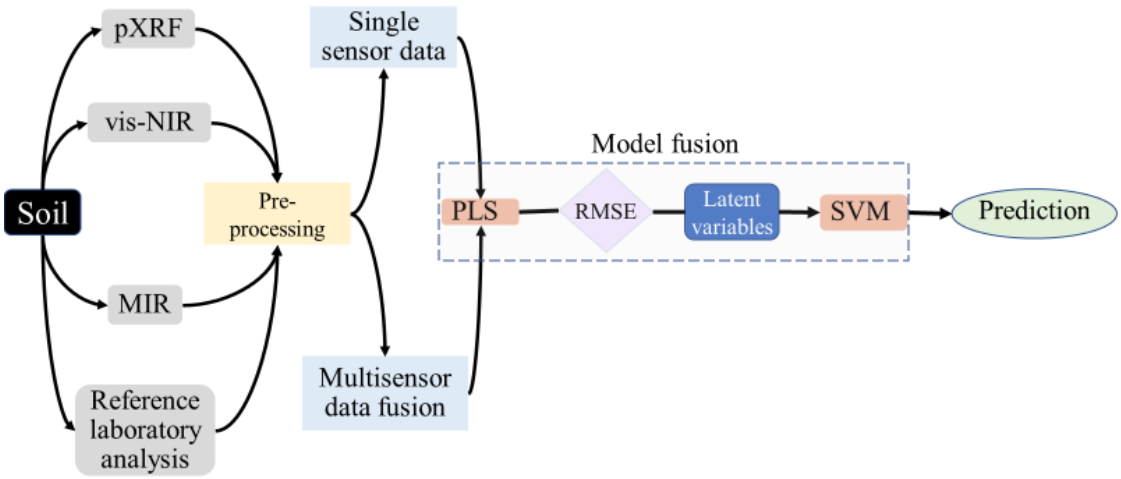


Fig. 3.2 A flow diagram of the model fusion process in the study.

3.2.5 Combined use of proximal sensors data

Proximal sensors' data were combined either through spectral data fusion (Section 2.4.3.3) or model averaging (Section 2.4.3.4). Spectral data fusion was done by concatenating the spectral data of the proximal sensors tested. Two methods of model averaging were tested, BMA (Clyde, 1999) as implemented in the R package *BAS* (Clyde, 2020) and GRA (Granger and Ramanathan, 1984) as implemented in the R package *GeomComb* (Weiss and Roetzer, 2016).

3.2.6 Evaluation of model performance

The predictive accuracy of constructed models was assessed and compared using several performance statistics: RMSE, R^2, RPIQ (Bellon-Maurel et al., 2010), CCC (Lin, 1989), and bias (Section 2.4.5). These parameters were calculated using the R package *spectacles* (Roudier, 2021). Statistical summaries and correlations were calculated and tabulated for the soil TE and spectrally active soil components using R packages *psych* (Revelle, 2021) and *Hmisc* (Harrell and DuPont, 2021).

3.3 Results

3.3.1 Concentrations of soil TE and spectrally active soil components

The total concentrations of soil As, Cd, Cr, Cu, Hg, Ni, Pb, Zn, and selected spectrally active soil components Al, C, and Fe are summarised in Table 3.1. The mean concentrations of TE were 4.86 mg As/kg, 0.08 mg Cd/kg, 13.81 mg Cr/kg, 13.68 mg Cu/kg, 0.05 mg Hg/kg, 8.95 mg Ni/kg, 12.93 mg Pb/kg, and 48.95 mg Zn/kg soil (Table 3.1). The TE concentrations were in general positively skewed, with Cd being the most skewed TE in the training set (skewness=2.35), and Cr the most skewed TE in the testing set (skewness=3.22; Table 3.1). Mean concentrations of spectrally active soil components were 16.52 g Al/kg, 42.41 g C/kg, and 21.37 g Fe/kg soil (Table 3.1).

Table 3.1 Statistical summary of total soil concentrations of TE (As, Cd, Cr, Cu, Hg, Ni, Pb, and Zn) and spectrally active soil components (Al, C, and Fe) in training and testing sets.

Parameters	As	Cd	Cr	Cu	Hg	Ni	Pb	Zn	Al	C	Fe
	mg/kg								g/kg		
Training set											
Number of samples	488	490	496	491	491	493	494	489	490	480	495
Minimum	0.30	0.01	3.50	1.24	0.00*	0.80	2.93	4.80	1.20	1.30	3.10
1st Quartile	2.70	0.03	9.40	7.97	0.03	5.60	9.97	32.80	12.13	28.38	16.40
Median	4.40	0.05	12.10	12.50	0.04	8.20	12.64	47.60	15.30	40.45	21.50
Mean	4.91	0.07	13.94	14.16	0.05	9.25	13.19	48.61	16.95	43.29	22.07
3rd Quartile	6.40	0.10	16.30	18.30	0.07	11.90	15.82	63.80	20.20	52.85	26.75
Maximum	16.40	0.56	44.50	55.16	0.20	31.20	34.28	105.20	52.80	118.30	49.00
Skewness	1.05	2.35	1.54	1.46	1.38	1.33	0.88	0.26	1.54	0.93	0.49
Testing set											
Number of samples	124	124	124	122	113	118	120	118	118	120	124
Minimum	1.00	0.01	3.10	1.56	<0.01*	1.00	4.39	6.50	4.40	9.30	5.10
1st Quartile	2.70	0.05	9.45	7.62	0.03	5.70	9.01	37.15	11.55	31.28	14.10
Median	3.95	0.07	11.85	11.10	0.05	7.30	11.43	48.50	13.85	37.40	17.95
Mean	4.67	0.09	13.30	11.73	0.05	7.73	11.88	50.35	14.71	38.88	18.58
3rd Quartile	5.60	0.12	15.32	15.05	0.07	9.95	14.18	61.15	17.23	46.30	22.23
Maximum	20.80	0.43	60.20	25.56	0.13	17.50	26.28	95.20	28.70	77.70	40.40
Skewness	2.24	1.92	3.22	0.58	0.63	0.54	0.94	0.23	0.78	0.32	0.57

*Minimum Hg concentration reported was 0.0025 mg Hg/kg soil.

The correlations among TE and spectrally active soil components (Al, C, and/or Fe) are shown in Table 3.2. Correlations were significant ($p<0.001$) and positive between Cr and Cu ($r=0.55$), Cr and Ni ($r=0.72$), Cu and Ni ($r=0.72$), Cu and Zn ($r=0.69$), and Ni and Zn ($r=0.70$) (Table 3.2). Among TE, Cd and Cr had a significant positive correlation ($p<0.01$, $r>0.10$) with Al, C, and Fe concentrations (Table 3.2). Arsenic, Ni, Pb, and Zn had a significant negative correlation ($p<0.05$, $r>-0.07$) with C concentrations (Table 3.2). Correlations between TE and spectrally active components were significant ($p<0.001$) positive for Cr and Cu with Al ($r>0.50$), Hg with C ($r=0.56$) and Cr, Cu, Ni, and Zn with Fe ($r>0.60$; Table 3.2).

Table 3.2 Correlation matrix among total soil concentrations of TE (As, Cd, Cr, Cu, Hg, Ni, Pb, and Zn) and spectrally active soil components (Al, C, and Fe) for the combined dataset used for training and testing.

	As	Cd	Cr	Cu	Hg	Ni	Pb	Zn	Al	C
As										
Cd	-0.05									
Cr	0.08	0.36***								
Cu	0.32***	0.26***	0.55***							
Hg	-0.23***	0.23***	0.15***	0.00						
Ni	0.34***	0.27***	0.72***	0.72***	-0.15***					
Pb	0.42***	-0.10*	0.03	0.33***	-0.15***	0.23***				
Zn	0.37***	0.39***	0.49***	0.69***	-0.11**	0.70***	0.37***			
Al	-0.01	0.34***	0.66***	0.56***	0.33***	0.44***	0.05	0.45***		
C	-0.25***	0.21***	0.13**	-0.07	0.56***	-0.13**	-0.14***	-0.08*	0.32***	
Fe	0.41***	0.19***	0.65***	0.77***	0.08	0.68***	0.31***	0.62***	0.73***	0.01

Level of significance: *p<0.05; **p<0.01; ***p<0.001

3.3.2 Performance of prediction models quantifying TE

In general, models using a particular combination of proximal sensors information via data fusion or model averaging outperformed those models obtained from a single sensor (Fig. 3.3; Appendices 3.2 and 3.4). This is evidenced by reduced RMSE and bias, and increased R^2, CCC, and RPIQ, suggesting improved accuracy of estimations for multiple TE (Fig. 3.3; Appendix 3.2). Among three proximal sensors, PLS-SVM based models using pXRF data performed fairly to quantify As, Cr, Cu, Hg, Pb, Ni, and Zn, whereas vis-NIR data performed satisfactorily predicting Cd with relatively low RMSE (Fig. 3.3; Appendices 3.2 and 3.4). Models using MIR data did not outperform other sensors to quantify TE (Appendix 3.2).

The PLS-SVM models using fused pXRF and MIR data performed well predicting As, Cr, and Pb concentrations (Fig. 3.3a; Appendix 3.2). These models showed CCC of 0.97 and RMSE of 0.57 mg As/kg, 1.14 mg Cr/kg, and 1.05 mg Pb/kg, respectively (Fig. 3.3a; Appendix 3.2). The PLS-SVM models, using fused pXRF and vis-NIR data, performed very well when predicting Cu, Ni, and Zn concentrations (CCC, 0.98; RMSE, 1.09 mg Cu/kg, 0.63 mg Ni/kg, and 3.90 mg Zn/kg, respectively; Fig. 3.3b; Appendix 3.2).

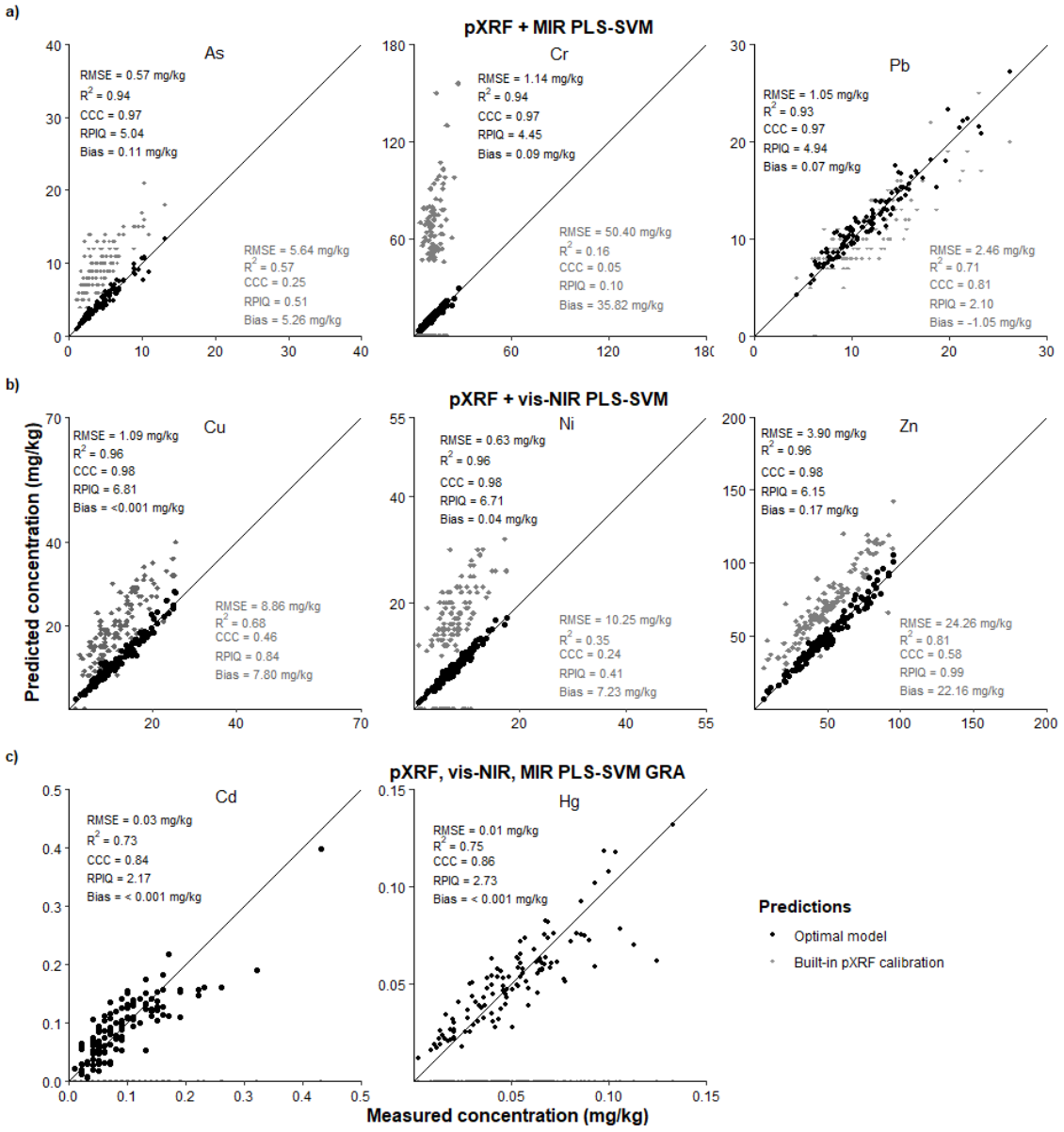


Fig. 3.3 Measured versus predicted concentrations of a) As, Cr, and Pb, b) Cu, Ni, and Zn, and c) Cd and Hg for the testing set based on the optimal model using a combination of proximal sensors data coupled with chemometric methods (black dots; performance statistics in black letters) and the built-in pXRF calibration (grey diamonds; performance statistics in grey letters). A black diagonal line going through the origin is a 1:1 line. Spectral data fusion of pXRF + MIR as input for PLS-SVM performed optimally for As, Cr, and Pb. Spectral data fusion of pXRF + vis-NIR as input for PLS-SVM performed optimally for Cu, Ni, and Zn. GRA of PLS-SVM model-based outputs of pXRF, vis-NIR, and MIR data performed optimally for Cd and Hg. Both Cd and Hg concentrations were below the lower detection limit of the pXRF built-in calibration.

For Cd and Hg, the best prediction results were obtained using GRA to average model outputs based on PLS-SVM regression of pXRF, vis-NIR, and MIR independently (Fig. 3.3c, Appendix 3.2). These optimal models for Cd and Hg performed well, with CCC of >0.80 and RMSE of 0.03 mg Cd/kg and 0.01 mg Hg/kg, respectively (Fig. 3.3c, Appendix 3.2). An examination of pXRF, vis-NIR, and MIR combination weights using GRA for Cd were 0.24, 0.99, and 0.27, respectively. For Hg, pXRF, vis-NIR, and MIR combination weights using GRA were 0.39, 0.40, and 0.22, respectively. When BMA was used instead of GRA, the performance of models estimating Cd and Hg was similar to or lower than the performance of GRA, for the same combinations (Appendix 3.2). Notably, Cd and Hg concentrations predicted are low (Table 3.1), which can explain the relatively inferior model performance compared to other TE (Appendix 3.2).

3.3.3 Comparison of optimal prediction models with the built-in pXRF calibration to quantify TE

The best models quantifying multiple TE outperformed the direct quantification from the built-in pXRF calibration (Fig. 3.3). The built-in pXRF calibration performed with poor accuracy (i.e., non-optimal RMSE, CCC, and RPIQ) for most TE except Pb (Fig. 3.3; Appendix 3.5). For Pb, the built-in pXRF calibration performed well (RMSE=2.46 mg Pb/kg), with relatively high CCC (0.81), and RPIQ (2.10; Fig. 3.3a; Appendix 3.5). For predicting As, Cr, Cu, Ni, and Zn, the built-in pXRF calibration showed poor accuracy with high RMSE, non-optimal CCC, and RPIQ (Figs 3.3a and 3.3b; Appendix 3.5). Built-in pXRF calibration did not quantify Cd and Hg due to their concentrations being below the detection limit of the pXRF instrument (5 mg Cd/kg, 1 mg Hg/kg; Fig. 3.3c; Appendix 3.5).

3.3.4 Important wavelengths for TE prediction

The PLS loadings highlighted the relative importance of certain spectral regions, and these areas contributed to explain the performance of a particular prediction model (Fig. 3.4). For As, Cr, and Pb, loadings showed the importance of the 3700–3000 cm^{-1} region (Al– and Fe–containing minerals) and 2130–1700 cm^{-1} region (metal–carbonyl (–CO)) of the MIR spectra (Fig. 3.4a; Appendix 2.3). For As, Cr, and Pb, the region of 6.40 keV

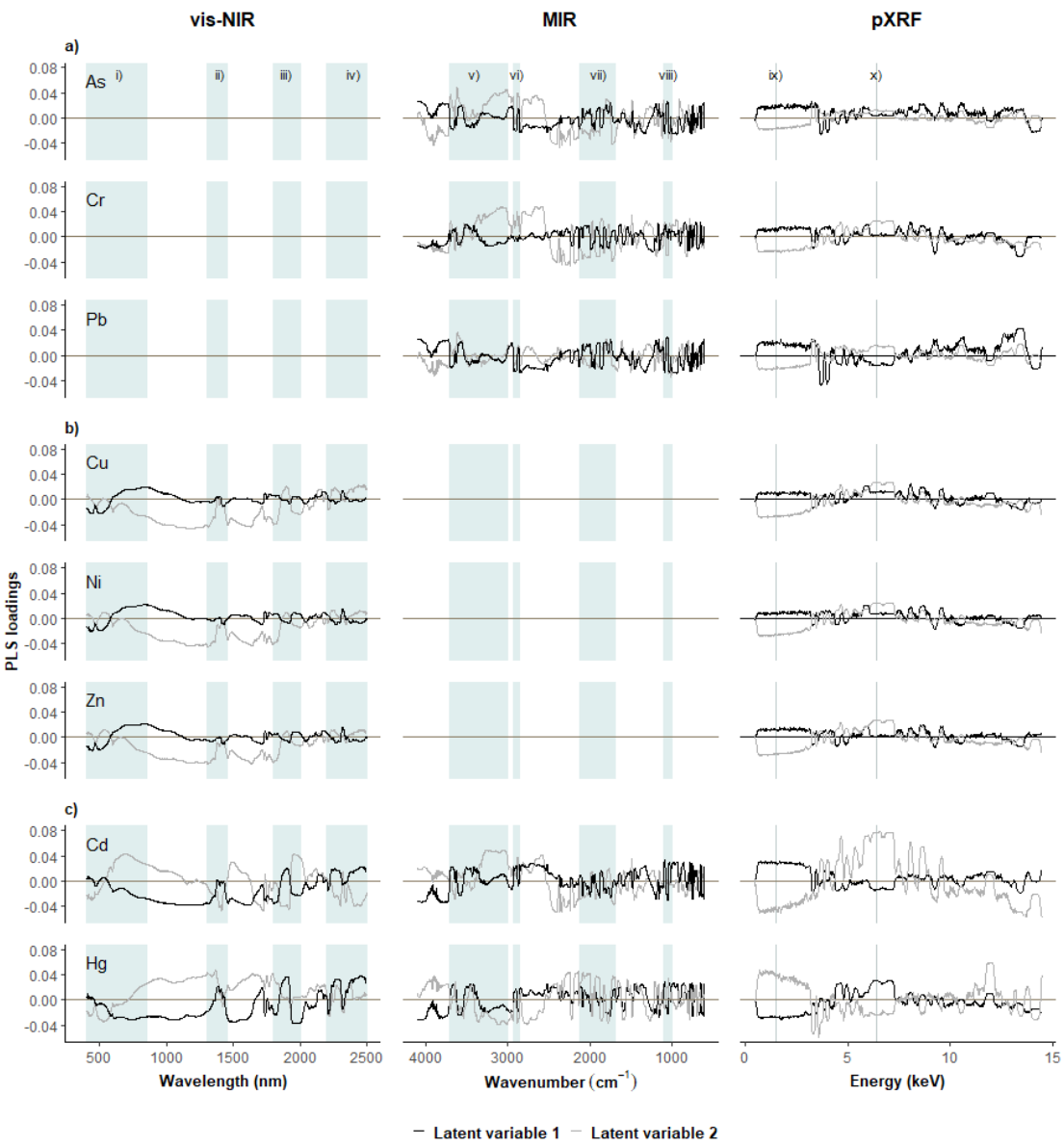


Fig. 3.4 PLS loadings for the first (black line) and second (grey line) latent variables in the optimal models using PLS-SVM model fusion with inputs from combinations of sensors: a) spectral data fusion of pXRF + MIR for As, Cr and Pb, b) spectral data fusion of pXRF + vis-NIR for Cu, Ni, and Zn, and c) GRA of PLS-SVM model outputs based on pXRF, vis-NIR, and MIR for Cd and Hg. The shaded areas in the vis-NIR region: i) soil colour, iron oxides, and soil organic matter (400–850 nm), ii) minerals (1300–1450 nm), iii) Al– and Fe–containing minerals and soil organic matter (1800–2000 nm), and iv) Al– and Fe–containing minerals and soil organic matter (2200–2500 nm); in the MIR region: v) Al– and Fe–containing minerals (3700–3000 cm^{-1}), vi) alkyl (2929–2855 cm^{-1}), vii) metal–carbonyl (2130–1700 cm^{-1}), and viii) quartz (1100–1000 cm^{-1}); and in the XRF region: ix) Al (1.48 keV) and x) Fe (6.40 keV).

(Fe) of XRF spectra was also important (Fig. 3.4a). For Cu, Ni, and Zn, several areas of vis-NIR spectra: 400–850 nm region (iron oxides and soil organic matter), 1300–1450 nm (Al– and Fe–containing minerals), and 1800–2000 nm (Al– and Fe–containing minerals and soil organic matter) as well as 6.40 keV (Fe) of XRF spectral region were highlighted as important (Fig. 3.4b; Table 2.6; Appendix 2.1). For Cd and Hg, a larger suite of wavelengths was used (Fig. 3.4c). These were in the visible (400–850 nm, iron oxides and soil organic matter), the near-infrared (1300–1450 nm, Al– and Fe–containing minerals; 1800–2000 nm, Al– and Fe–containing minerals and soil organic matter), the mid-infrared (3700–3000 cm^{-1}, Al– and Fe–containing minerals and soil organic matter; 2929–2885 cm^{-1}, alkyl ($–CH_3$); 2130–1700 cm^{-1}, metal–carbonyl (–CO)), and the XRF (1.48 keV, Al and 6.40 keV, Fe) regions of the electromagnetic spectrum (Fig. 3.4c; Table 2.6; Appendices 2.1 and 2.3).

3.4 Discussion

3.4.1 Performance of the built-in pXRF calibration to quantify TE

The built-in internal calibration of the pXRF instrument performed well for Pb, but poorly for the other TE (Fig. 3.3, Appendix 3.5). These results are in agreement with those of O'Rourke et al. (2016a), who also found fair accuracy in predicted Pb concentration, but poor accuracy for As, Cr, and Cu using the built-in pXRF calibration in Irish soils. For Zn, all soil concentration values were above the limit of detection by >3 mg Zn/kg; however, the built-in pXRF calibration still overestimated Zn concentration (mean of 72.51 mg Zn/kg versus 50.35 mg Zn/kg) than reference laboratory results (Fig. 3.3; Table 3.1; Appendix 3.5). This result is in agreement with Schneider et al. (2016) and O'Rourke et al. (2016a). Built-in pXRF calibration can successfully quantify some TE (e.g., Pb), whereas when measuring other TE at low concentrations, the pXRF should be recalibrated with local samples to yield accurate results.

3.4.2 Comparison of performance of optimal TE prediction models with published studies

These results demonstrate the feasibility of predicting As, Cr, and Pb by fusing pXRF and MIR data (Fig. 3.3a). Other studies (O'Rourke et al., 2016b; Xu et al., 2020) have been less successful in assessing these TE concentrations in soil either using data fusion or other computationally intensive methods such as outer product analysis or Cubist (Table 3.3). Previous studies have also used GRA model averaging to determine As, Cr, and/or Pb concentrations using pXRF, vis-NIR, and/or MIR with Cubist models performing poorly (Table 3.3; O'Rourke et al. (2016a); Pozza et al. (2020)). In contrast to the optimal models constructed for Cu, Ni, and Zn concentrations through fusing pXRF and vis-NIR data reported in this study (Fig. 3.3b), past attempts at data fusion and random forest regression of total XRF and MIR, have quantified extractable Cu and Zn with poor accuracy (Table 3.3; Towett et al. (2015b)).

3.4.3 Model fusion and model averaging for TE predictions

This study demonstrates that PLS-SVM is an effective approach to predict TE using the proximal sensors tested here, particularly when applying data fusion or model averaging (Fig. 3.3; Appendix 3.2). The PLS-SVM has been previously applied to predict a range of soil properties, including available water-holding capacity and soil organic carbon content using vis-NIR data only (Blaschek et al., 2019; Hong et al., 2019). The PLS-SVM model fusion combines features of PLS and SVM for improved TE predictions. The PLS algorithm performs matrix decomposition on the spectral and TE concentration data to extract latent variables (Bao et al., 2017) and this helps to (1) reduce data redundancy (Hong et al., 2019) and (2) remove noise (Blaschek et al., 2019) and matrix interference caused mainly by soil organic matter in XRF spectra (Ravansari and Lemke, 2018). The use of SVM regression on PLS latent variables takes into account the non-linear spectral association between TE and spectrally active soil components (Blaschek et al., 2019).

Table 3.3 Comparison of optimal TE prediction models from this study with the selected previous studies.

Proximal sensing techniques	Chemometric methods	TE	Min.	Mean	Max.	Area	No. of samples	Skewness	RMSE	R^2	CCC	RPIQ	References
			mg/kg			km^2			mg/kg				
pXRF + vis-NIR	OPA Cubist	As	0.20	12.45	37.37	1831	301	1.13	3.51	–	0.82	2.70	(Xu et al., 2020)
pXRF, vis-NIR, MIR	Cubist GRA		0.70	10.10	51.60	84,000	322	2.50	3.68	0.53	–	0.59	(O'Rourke et al., 2016a)
pXRF + MIR	**PLS-SVM**		**0.30**	**4.86**	**20.80**	**40,000**	**612**	**1.60**	**0.57**	**0.94**	**0.97**	**5.04**	**This study**
pXRF + vis-NIR	OPA Cubist	Cd	0.09	0.42	2.35	1831	301	2.70	0.15	–	0.82	1.39	(Xu et al., 2020)
pXRF, vis-NIR, MIR	Cubist GRA		0.01	0.60	5.00	84,000	322	2.70	0.41	0.35	–	0.56	(O'Rourke et al., 2016a)
pXRF, vis-NIR, MIR	**PLS-SVM GRA**		**0.01**	**0.08**	**0.56**	**40,000**	**614**	**1.29**	**0.03**	**0.73**	**0.84**	**2.17**	**This study**
pXRF + vis-NIR	Cubist	Cr	9.87	43.54	95.62	1831	301	0.00	8.04	–	0.86	2.97	(O'Rourke et al., 2016a)
pXRF, vis-NIR	Cubist GRA		6.40	49.48	221.70	84,000	322	–	10.33	0.77	0.86	1.02	(O'Rourke et al., 2016b)
pXRF + MIR	**PLS-SVM**		**3.10**	**13.86**	**60.20**	**40,000**	**620**	**2.25**	**1.14**	**0.94**	**0.97**	**4.45**	**This study**
TXRF + MIR	RF	Cu	0.001	2.00	24.00	100	700	–	2.10	0.57	–	1.09	(Towett et al., 2015b)
pXRF, vis-NIR	Cubist GRA		2.90	23.15	85.70	84,000	322	–	6.23	0.85	0.91	1.19	(O'Rourke et al., 2016b))
pXRF + vis-NIR	**PLS-SVM**		**1.24**	**13.68**	**55.16**	**40,000**	**613**	**1.97**	**1.09**	**0.96**	**0.98**	**6.81**	**This study**
pXRF, vis-NIR	Cubist GRA	Ni	4.38	20.76	51.97	1831	301	0.38	3.72	–	0.89	3.42	(Xu et al., 2020)
pXRF, vis-NIR, MIR	Cubist GRA		1.70	24.30	176.00	84,000	322	2.70	9.76	0.80	–	0.65	(O'Rourke et al., 2016a)
pXRF + vis-NIR	**PLS-SVM**		**0.80**	**8.95**	**31.20**	**40,000**	**611**	**1.40**	**0.63**	**0.96**	**0.98**	**6.71**	**This study**
pXRF, vis-NIR	Cubist GRA	Pb	3.20	143.20	2305.00	500	405	4.70	86.40	–	0.95	0.37	(Pozza et al., 2020)
pXRF + vis-NIR	OPA Cubist		0.78	23.35	169.45	1831	301	3.90	11.09	–	0.72	1.20	(Xu et al., 2020)
pXRF, vis-NIR, MIR	Cubist GRA		5.20	29.80	123.00	84,000	322	1.90	11.19	0.59	–	0.52	(O'Rourke et al., 2016a)
pXRF + MIR	**PLS-SVM**		**2.93**	**12.93**	**34.28**	**40,000**	**614**	**0.92**	**1.05**	**0.93**	**0.97**	**4.94**	**This study**
TXRF + MIR	RF	Zn	0.001	2.00	37.00	100	700	–	4.30	0.08	–	0.28	(Towett et al., 2015b)
pXRF, vis-NIR	Cubist GRA		16.00	76.63	240.60	84,000	322	–	19.73	0.73	0.84	0.94	(O'Rourke et al., 2016b))
pXRF + vis-NIR	**PLS-SVM**		**4.80**	**48.95**	**105.20**	**40,000**	**607**	**0.25**	**3.90**	**0.96**	**0.98**	**6.15**	**This study**

Taken individually, pXRF spectra as input for PLS-SVM outperformed both vis-NIR and MIR to quantify As, Cr, Pb, Cu, Ni, and Zn, due to the direct relation of pXRF spectral response with TE concentrations above the limit of detection (Appendix 3.2; Padilla et al. (2019)). For Cd with concentrations below the limit of detection, using vis-NIR spectra outperformed pXRF as input for PLS-SVM (Appendices 3.2 and 3.3). The pXRF predictions below the detection limit deviated from the measured values due to the overlapping of low-intensity L-line spectral response of Cd by strong K-line spectral response of the soil matrix mineral component containing K and Ca (Appendix 3.3; Elam et al. (2002); Stosnach (2005)).

The improved performance of vis-NIR in Cd quantification is based on the indirect association of this TE with spectrally active soil components involved in its retention, i.e., soil organic matter and Al– and Fe–containing minerals (Kooistra et al., 2001), as indicated by (1) PLS loadings (Fig. 3.4c) and (2) a significant correlation between Cd and Al, C, and Fe concentrations (Table 3.2). However, the lack of a consistent association of Hg with Fe minerals (Fig. 3.4c; Table 3.2) limited the accuracy of models using vis-NIR data to quantify this element (Appendix 3.2).

Models based on spectral data fusion of pXRF with vis-NIR or MIR reflectance spectra as input for PLS-SVM outperformed pXRF-based models alone for quantifying As, Cr, Pb, Cu, Ni, and Zn (Appendices 3.2 and 3.4). Models based on pXRF spectra performed fairly with some deviations from the measured concentrations due to spectral interferences caused by soil matrix components as organic matter and minerals containing Mg, Ca, Al, Fe, and Mn (Elam et al., 2002; Ravansari and Lemke, 2018; Stosnach, 2005). Both organic matter and minerals (1) contribute to vis-NIR and/or MIR reflectance spectra and (2) regulate TE binding in soils (Kooistra et al., 2001), revealing indirectly associated TE (Fig. 3.4; Wu et al. (2010)), and thus TE quantification was more accurate and precise when reflectance spectra were combined with pXRF spectra (Appendices 3.2 and 3.4).

Using PLS-SVM with the fused pXRF and MIR dataset emphasised the association of TE (e.g., As, Cr, and Pb) with Fe containing minerals (Fig. 3.4a; Appendix 2.3; Janik et al. (1998); Li et al. (2021); Niazi et al. (2015); Wang et al. (2017)). These associations were based on the correlation of As, Cr, and Pb concentrations with Fe concentration (Table 3.2). Similarly, applying PLS-SVM to the fused spectral dataset of pXRF and vis-NIR highlighted the associations of Cu, Ni, and Zn with Fe– and Al–containing minerals in specific regions of the vis-NIR spectra (Fig. 3.4b; Appendix 2.1; Ben-Dor (2002); Soriano-Disla et al. (2014);

Stenberg and Rossel (2010); Wang et al. (2015)). In addition, Cu, Ni, and Zn concentrations and Fe– and Al–concentrations were correlated (Table 3.2). Due to the ability to capture associations between TE and spectrally active soil components (Wu et al., 2010) found in nature (Martin et al., 2016), PLS-SVM succeeded in quantifying TE at low concentrations (Fig. 3.3).

Although both model averaging methods (GRA and BMA) performed with similar accuracy (Appendix 3.2), the simplicity of GRA method to implement and execute, and the wider history of use in soil TE quantification (Table 2.13; O'Rourke et al. (2016b); Pozza et al. (2020); Xu et al. (2020)), may prioritise the selection of GRA over BMA (Arsenault et al., 2015). Granger-Ramanathan model averaging method improved the estimation of Cd and Hg (Appendix 3.2), assigning a higher combination weight to vis-NIR model output. This is explained by PLS loadings which highlight the association of Cd and Hg with soil organic matter and Al–containing minerals, and the association of Cd with Fe oxides and Al– and Fe–containing minerals in the vis-NIR spectra (Fig. 3.4c; Appendix 2.1; Kemper and Sommer (2002)). Furthermore, Cd and Hg concentrations were correlated to Al and C concentrations, and Cd concentration was also correlated to Fe concentration (Table 3.2). For samples with TE concentrations well below the pXRF detection limit, the XRF spectral response alone often does not suffice to accurately quantify TE (Benedet et al., 2020; Javadi et al., 2021). This instance may require adding information from vis-NIR and MIR spectra via model averaging (O'Rourke et al., 2016a). Through capturing these associations, GRA averaging of PLS-SVM based model outputs of each proximal sensor (pXRF, vis-NIR, and MIR) used to assess Cd in this research could be usefully deployed for detailed monitoring of soil Cd concentration on individual farms to support policy such as the cadmium management strategy for sustainable pasture management in New Zealand (CWG, 2011). The level of precision and accuracy reported in this study makes GRA approach a good candidate for quantifying the Cd concentration of New Zealand agricultural soils.

3.4.4 Broader implications

The soil dataset used in this study was collected across a large part of the South Island in New Zealand, where pedological diversity is high (Fig. 3.1; Appendix 3.1; Hewitt et al. (2021)). The dataset spans a broad range of soil orders, geology, climatic conditions, land uses, and soil TE concentrations (Table 3.1; Appendix 3.1). Despite this variability, the methods developed in

this study were able to successfully replicate the TE concentrations in the samples analysed by reference laboratory method (Appendix 3.2; Gazley et al. (2020); Martin et al. (2016)). This indicates, with strong confidence, that the prediction models from this study can be successfully applied in other regions of New Zealand and would also be applicable internationally. The systematic combination approach taken in this study could be further used to (1) reduce the economic costs of sample preparation and analytics, (2) facilitate detailed monitoring of TE concentrations on individual farms for sustainable soil management, as well as 3) extend the monitoring beyond total TE concentrations to include an assessment of TE fractions associated with soil components, such as soil organic matter and Al– and Fe–containing minerals (Chapter 4; Chakraborty et al. (2017); Cipullo et al. (2019)). This advancement in soil testing could also be applied to other aspects of soil management (mineral composition, soil carbon storage, soil fertility management) because the spectral information gathered can be used to predict many other soil properties after proper validation (Nocita et al., 2015; Soriano-Disla et al., 2014; Zhang and Hartemink, 2020).

With the portability of pXRF, vis-NIR, and MIR sensors, these proximal sensors can be directly deployed in field conditions including airborne remote sensing (Chakraborty et al., 2022). However, prediction models developed in this study were based on laboratory conditions (sample drying, sieving, homogenisation, and/or grinding). Edaphic and environmental variables such as soil moisture content and particle size can influence field spectral measurements (Lemière, 2018; Roudier et al., 2017). Hence, further work is required to validate these models for the application of these proximal sensors in field situations.

3.5 Conclusions

This study successfully quantified multiple TE concentrations in soil (As, Cd, Cr, Cu, Hg, Ni, Pb, and Zn) using different combinations of proximal sensors (pXRF, vis-NIR, and MIR). The rich dataset was then tested using a compressive suite of modelling approaches including PLS versus PLS-SVM model fusion, and single sensor predictions versus multi-sensor data fusion versus single sensor predictions averaging. The model fusion approach (PLS-SVM) systematically outperformed PLS regression, by considering non-linearities in the relationships between sensor data and variations in TE concentrations. For As, Cr, and Pb, the best predicted concentrations were obtained using PLS-SVM fused pXRF and MIR data. For Cu, Ni, and Zn, the best prediction results were obtained through fusing pXRF and vis-NIR data. For Cd and

Hg, GRA model averaging of PLS-SVM based model outputs of individual pXRF, vis-NIR, and MIR, performed the best although performance was not as good as for other TE. Additionally, the best model for each TE in this study systematically outperformed the built-in pXRF calibration results. The combination of chemometric methods resulted in (1) a better lower method of the detection limit of the pXRF alone, (2) reduced organic matter interference in the pXRF spectra, and (3) highlights the association between TE and spectrally active soil components (e.g., Al– and Fe–containing minerals) not evident from pXRF alone. Overall, the prediction models fitted for the different TE assessed performed with reliable accuracy. These methods can be applied for cost-effective TE analysis of soil samples collected from regional- to national-scale geochemical baseline surveys and for monitoring and reporting studies on specific TE concentrations, such as the New Zealand Cadmium Management Strategy.

Chapter 4

Predicting total cadmium and cadmium fractions in agricultural soils using vis-NIR, MIR, and pXRF spectroscopy

Abstract

Long-term continuous application of phosphate fertiliser to sustain plant production has unintentionally accumulated Cd in some New Zealand agricultural soils. Monitoring of total soil Cd is recommended for effective management of soil Cd and to minimise plant accumulation. Compared to total Cd, analysing chemical forms by sequential extraction can provide information on soil Cd availability, distribution, and mobility to assess the potential plant Cd uptake accurately. A total of 87 air-dried topsoil (0–15 cm) samples from long-term pastoral farms with phosphate fertiliser application history in New Zealand were analysed by reference laboratory methods for total Cd and Cd distribution in exchangeable, acid soluble, metal oxides bound, organic matter bound, and residual fractions. The data acquired using three proximal sensing techniques, vis-NIR, MIR, and pXRF spectroscopy were used as input for PLS regression to develop models predicting total Cd and Cd fractions. The average total soil Cd concentration was 0.58 mg Cd/kg soil. For total Cd, cross-validation (cv) results of models using individual vis-NIR, MIR, and pXRF data performed with $nRMSE_{cv}$ of 26%, 30%, and 31% and CCC_{cv} of 0.85, 0.77, and 0.75, respectively. For acid soluble and organic matter bound Cd, models using vis-NIR data performed with $nRMSE_{cv}$ of 11% and 33% and CCC_{cv} of 0.97 and 0.84, respectively. For exchangeable, metal oxides bound, and residual Cd, models using MIR data performed with $nRMSE_{cv}$ of 40%, 67%, and 100% and CCC_{cv} of 0.57, 0.61, and 0.59, respectively. Proximal sensing techniques could potentially be applied to quantify total Cd and potential plant-available Cd fractions for effective implementation of Cd monitoring programmes such as the TFMS in New Zealand.

4.1 Introduction

Agricultural intensification has raised concerns on soil health management in New Zealand and globally (Drewry et al., 2021; Lehmann et al., 2020; MacLeod and Moller, 2006). Long-term continuous phosphate fertiliser application to sustain plant production during the last century, unintentionally increased Cd concentration in agricultural soils as this fertiliser contained traces of Cd (Abraham, 2018). Elevated concentration of soil Cd may enhance Cd availability, leading to uptake and accumulation of Cd in plant edible parts (Stafford et al., 2016). Consequently, year-round pastoral grazing can expose animals (e.g., sheep, dairy cattle) to long-term dietary accumulation of Cd (Roberts et al., 1994). Such exposure can adversely affect animal food product quality, resulting in dietary and trade risks when provisional safe limits for Cd uptake are exceeded (ATSDR, 2012; Clemens et al., 2013; McDowell and Gray, 2022).

To minimise risks and manage soil Cd accumulation in New Zealand agricultural soils, a national Cadmium Management Strategy was developed in 2011 with the subsequent implementation of a TFMS (CWG, 2011). The TFMS recommends total soil Cd testing in an accredited laboratory repeated every five years to monitor the long-term trend in Cd concentration and mitigate its flow-on effect on the food chain. The TFMS sets out increasingly stringent restrictions on the choice and rate of phosphate fertiliser application as total soil Cd concentration increases (Fertiliser Association, 2019).

Total concentration analysis may not be enough to assess environmental risk, plant availability, and mobility of Cd (Kabata-Pendias, 2010). Only a portion of total soil Cd may be of immediate harm to the environment (Yi et al., 2020). In soil solution, Cd occurs as a free exchangeable divalent cation or can form soluble complexes with Cl and S or associate with dissolved organic matter (Kubier et al., 2019). Cadmium can also be sorbed onto the soil solid phase by: (1) interacting with carboxylic and phenolic functional groups forming organic ligand–Cd^{+}, organic ligand–Cd^{-} chemical structures with soil organic matter (Loganathan et al., 2012; Traina, 1999), (2) forming stable complexes with clays (Al– and Fe–containing minerals), or (3) precipitating as $CaCdCO_3$ and $Cd_3(PO4)_2$ compounds (Bolan et al., 2003b) (Detailed description in Section 2.1.5).

Characterising soil Cd distribution among geochemical fractions (e.g., exchangeable, acid soluble, or organic matter bound Cd) is critical to assess available Cd and subsequent

plant Cd uptake (Krishnamurti, 2008). Exchangeable Cd is defined as weakly sorbed Cd pool in the soil, including both water-soluble and exchangeable species (Tessier et al., 1979). The amount of exchangeable Cd is generally used as the proxy for phyto-available Cd in soil (Gleyzes et al., 2002). The Cd concentration in the exchangeable fraction is controlled by the sorption and desorption of Cd in other soil fractions (Loganathan et al., 2012). As a consequence of changes in sorption/desorption, Cd may be liberated for leaching or uptake in plants, depending on soil properties such as pH, soil organic matter, the abundance of Al– and Fe–containing minerals, and cation exchange capacity (Gray et al., 2019b; Naidu et al., 1994; Stafford et al., 2018c; Yuan and Lavkulich, 1997). Hence, along with exchangeable fraction, acid soluble and organic matter bound Cd fractions may contribute to the plant Cd uptake (Chavez et al., 2016; El-Mufleh Al Husseini et al., 2013).

The chemical forms of Cd in the soil solid phase can be studied using sequential extraction methods, which recover target elements from geochemical fractions (Adriano, 1992). These methods broadly separate Cd into one of five operationally-defined fractions: exchangeable, acid soluble, metal oxides bound, soil organic matter bound, and residual Cd (Table 2.3; Section 2.2.2; Gleyzes et al. (2002)). Sequential extraction provides information about the origin, mode of occurrence, mobility, and availability of elements in the soil and this information is essential to understand plant uptake (Kabata-Pendias, 2010; Krishnamurti, 2008). However, the recovery of total Cd, when compared to total extraction, is sometimes less than 100% (Bacon and Davidson, 2008).

Proximal sensing techniques can potentially be a fast and cost-effective alternative to reference laboratory approaches (Nawar et al., 2019; Shepherd and Walsh, 2002). Proximal sensing techniques, such as vis-NIR (350–2500 nm), MIR (2500–25000 cm^{-1}), and pXRF (0–40 keV) spectroscopy, have been assessed to quantify total Cd in contaminated soils (Kemper and Sommer, 2002; Padilla et al., 2019; Reeves III and Smith, 2009). For wider implementation of proximal sensors to quantify low total Cd concentration and Cd fractions in agricultural soils requires more extensive studies (Song et al., 2012; Stafford et al., 2018b; Wang et al., 2017). Stafford et al. (2018b) estimated site specific total soil Cd based on the significant relationship between Cd and model based on vis-NIR predicted soil total C and total N concentration. Similarly, vis-NIR has been used to assess exchangeable Zn and Pb in forest soils (Chodak et al., 2007) and

potentially-available Pb in contaminated agricultural soils (Paltseva et al., 2022). In other studies, vis-NIR has also been used to assess five As fractions in landfill site soils (Chakraborty et al., 2017), and the sum of exchangeable and acid-soluble fractions of As, Cd, Cr, Cu, Ni, Pb, and Zn in contaminated soils (Cipullo et al., 2019). These studies have used reflectance spectra (350–2500 nm) capturing response from spectrally active soil components (e.g., organic matter and Al–, Fe–, and Mn–containing minerals) with which Cd is bound (Kooistra et al., 2001; Wang et al., 2017). Other studies have used fluorescence spectra to capture response from elements in the matrix to directly estimate Cd and other TE concentrations (O'Rourke et al., 2016a). However, these sensors have not been examined to quantify low total Cd concentration and Cd fractions in New Zealand agricultural soils.

The applicability of proximal sensing techniques to assess total soil Cd and Cd fractions concentrations can contribute to the cost-effective monitoring of Cd and minimise plant Cd accumulation in agricultural systems (Nocita et al., 2015). The potential of using proximal sensors to quantify low total Cd concentration and Cd fractions in agricultural soils requires more extensive studies for wider implementation (Cipullo et al., 2019; Song et al., 2012; Stafford et al., 2018b). However, these proximal sensors have been studied one sensor at a time, mostly under contaminated soils using limited samples and coverage. In the present study, agricultural soils exposed to long-term phosphate fertiliser application history from pastoral farms in New Zealand were assessed for total Cd and Cd fractions. This assessment was accomplished following (1) reference laboratory analysis and (2) proximal sensing techniques including individual vis-NIR, MIR, and pXRF sensors.

4.2 Materials and methods

4.2.1 Soil sampling and chemical analyses

Soil samples were received from the previous study by Stafford (2017). Briefly, topsoil (0–15 cm) samples (n=87) from 30 commercial pastoral farms with long-term phosphate fertiliser application history were collected from the North Island (22) and South Island (8) of New Zealand (Fig. 4.1). Collected soils were from dairy (23) and sheep and beef farms (7) that were exposed to long-term phosphate fertiliser application. The dominant

pasture cover of the sampled farms was chicory or plantain (56%), while the remaining pastures were under ryegrass/white clover. Soil samples covered soil orders representing the major agricultural soils in New Zealand as per NZSC, including Allophanic (57%), Pumice (16%), Organic (11%), Pallic (7%), Brown (5%), Gley (2%), and Recent (1%) soils. A composite sample for each site in the farm was prepared by aggregating soil cores from 15 subsampling locations (Stafford, 2017).

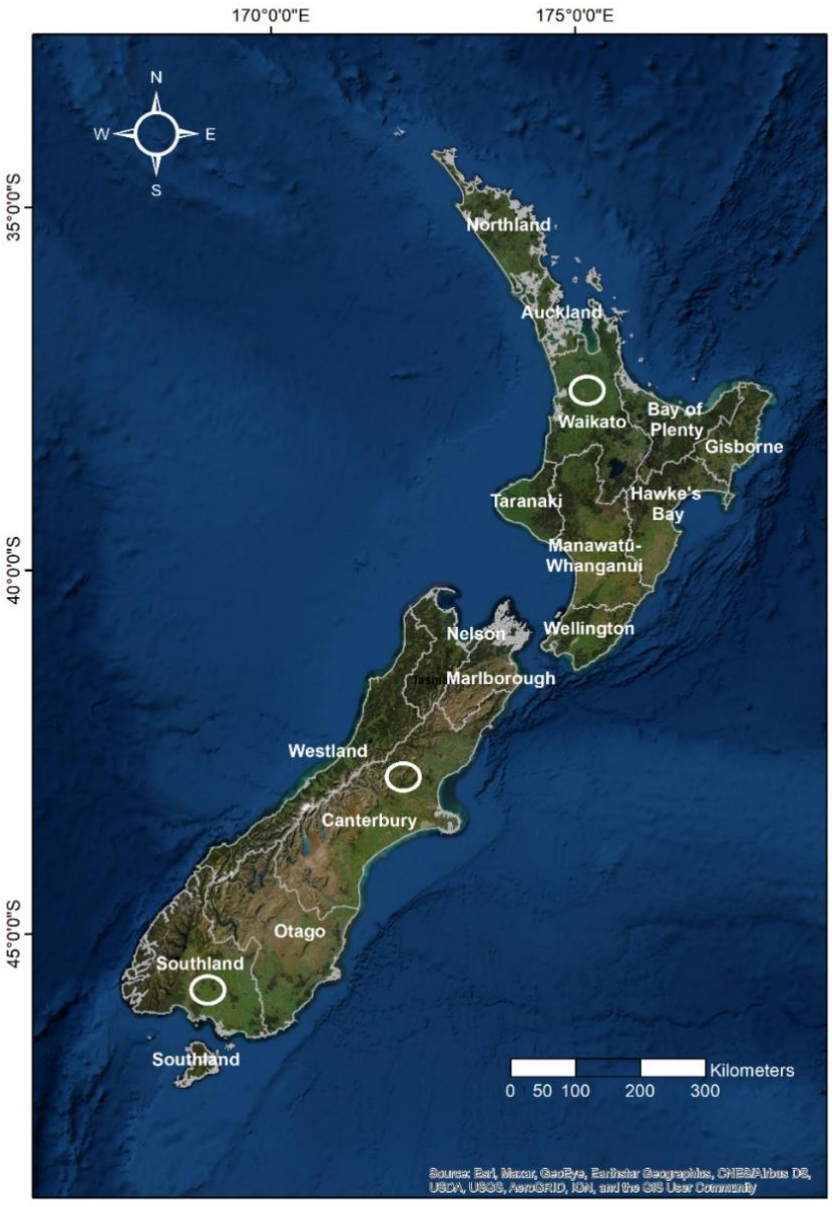


Fig. 4.1 A map of New Zealand outlining the general location of soil sampling areas in the North Island (Waikato region) and the South Island (Canterbury and Southland regions).

Soil samples were air-dried and sieved through a 2 mm sieve. Soil pH was measured at a ratio of 10 g soil to 25 ml deionised water using a Hanna Instruments HI 2211 pH/ORP meter. Acid oxalate extractable aluminium (Al_o), iron (Fe_o) and silicon (Si_o) were extracted using 0.40 g soil in 40 ml acid oxalate solution at pH 3 mixed by end-over-end shaker for 4 h (Blakemore et al., 1987) then analysed by microwave-plasma atomic emission spectrometer (MP-AES 4200, Agilent, USA). Total phosphorus (P) was extracted digesting 1.0 g soil in the 4 ml mixture of concentrated sulphuric acid, potassium sulphate, and selenium at 350°C for 4 h and analysed by autoanalyser using nitric-vanadomolybdate acid colouring reagent (McKenzie and Wallace, 1954). Cation exchange capacity (CEC) was determined using 1.0 g soil mixed with 3.0 g sand for solid phase extraction with 1 M ammonium acetate (at pH 7) (Blakemore et al., 1987) and analysed by MP-AES for cations: sodium, magnesium, potassium, and calcium. Soil total C and N were determined after adding an equal amount of tungsten oxide to the sub-0.5 mm 100 mg subsample, igniting it in the induction furnace and analysing it by a Vario MACRO cube elemental system (Elementar Analysensysteme GmbH, Hanau, Germany).

4.2.2 Soil cadmium analysis

4.2.2.1 Total Cd

Total soil Cd was determined from a sub-2 mm 1 g subsample with 10 ml concentrated nitric acid (HNO_3) left overnight and then digested for 2 h at 120°C in the heating block (Kovács et al., 2000). The remaining solution was made up to 25 ml volume with 2% HNO_3 before analysing by the graphite furnace atomic absorption spectrometer (GFAAS PinAAcle 900Z, PerkinElmer, UK). For total Cd analysis, the certified reference material CRM 051-050 clay 2 sample with a concentration of 42.2 mg Cd/kg soil was used. The measured mean concentration of CRM 051-050 clay 2 was 43.5 mg Cd/kg with a range of 94–106% of the expected value.

4.2.2.2 Sequential extraction of Cd

The sequential extraction procedure of Tessier et al. (1979) was used in this work, with modifications. The procedure separates Cd into five operationally defined fractions: exchangeable, acid soluble, metal oxides bound, organic matter bound, and residual Cd (Fig. 4.2; Table 2.3).

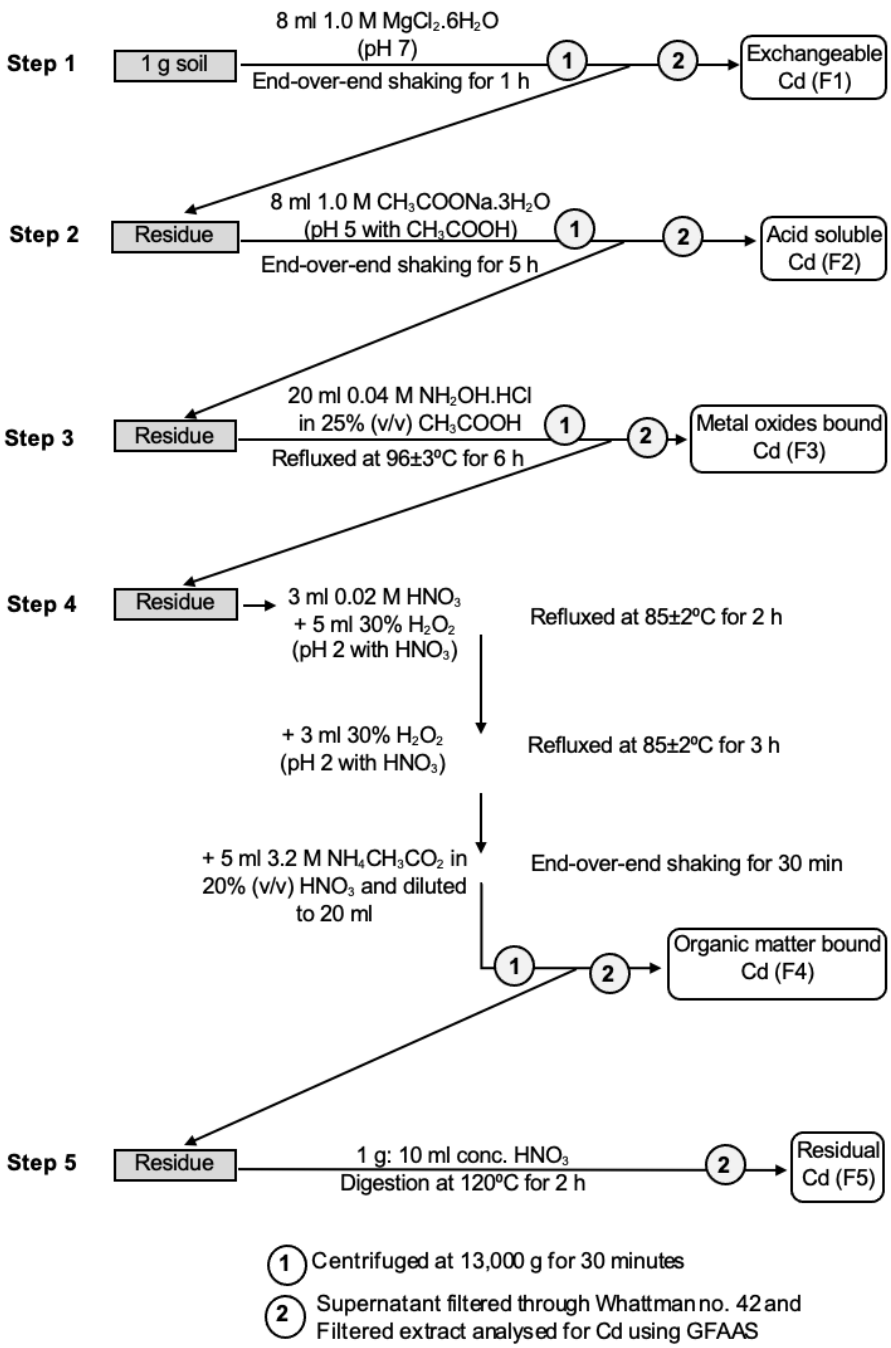


Fig. 4.2 Steps of sequential extraction of soil Cd fractions following Tessier et al. (1979) with modifications.

After each extraction, the sample was centrifuged at 13,000 g for 30 minutes and the supernatant was filtered through Whattman no. 42 to obtain a solution for further analysis

(Fig. 4.2). Residue from the previous step was washed using 8 ml of deionised water to remove any solution left before using it for the step. Filtered extracts were analysed for Cd by GFAAS.

4.2.3 Sample scanning, spectra collection, and pre-processing

Air-dried sub-2 mm representative soil subsamples were scanned using three proximal sensing techniques: vis-NIR, MIR, and pXRF. For vis-NIR scanning, soil samples were filled in a petri dish to a 10 mm depth (Viscarra Rossel et al., 2016). For MIR, sub-0.5 mm subsamples (four replicates per sample) were placed in an aluminium microtiter plate with 48 wells and vertical gutters in-between lines of 8 wells including four replicates per sample. For pXRF, a cylindrical plastic cup 22 mm high by 14 mm in diameter was filled with air-dried, sub-2 mm soil and covered with a polypropylene transparent film of 4 µm thickness.

Details on sensors and scanning methodology can be found in Chapters 2 (Table 2.7; Fig. 2.4) and 3. Briefly, an ASD FieldSpec 3 spectroradiometer was used to record a vis-NIR (350–2500 nm) reflectance spectra which was an average of 50 internal readings within the scanning period. Each sample was scanned in triplicate, rotating clockwise 90 degree angle. The ASD probe was pushed in contact with samples, ensuring no external light interferes with the measurement. A Fourier-transform infrared spectrometer (Vertex 70, Bruker, Germany) equipped with a microplate reader extension for high throughput screening infrared spectroscopy equipment (HTS-XT, Tensor II, Bruker, Germany) was used to record diffuse MIR (7498–600 cm^{-1}) reflectance spectra of samples (four replicates per sample). An Olympus Vanta C series pXRF instrument containing a rhodium anode was used to measure XRF spectra (0–40 keV) in *Geochem (2 beams)* mode using settings of 45 seconds for each beam and each sample was scanned twice.

The Mahalanobis distance was used to remove outliers from the elemental dataset (Section 2.4.2). A sample having a Mahalanobis distance score >1 was considered an outlier. Three samples containing higher Cd concentration (1.59, 1.75, and 2.03 mg Cd/kg soil) were outliers and removed from the dataset before prediction model development. These steps were performed using the R package *prospectr* (Stevens and Ramirez Lopez, 2020).

The principal component analysis (PCA) of spectral data was performed to reduce the multi-dimensionality of spectral data into a few principal components. Spectral data were centred before performing PCA. The distribution of Allophanic and non-Allophanic soil samples in the principal components space were plotted using the first two principal components describing the greatest amount of variance in the dataset.

4.2.4 Modelling framework

The R statistical environment (R Core Team, 2021) was used for all chemometric analysis. The spectral data (predictor variables) and reference laboratory Cd analysis results (target property) were used to cross-validate (which excludes the need to maintain separate calibration and validation sets) the individual sensor performance (Chodak et al., 2007). All concentration values were log-transformed and then transformed using mean centring and variance scaling (Section 2.4.1; Table 2.8) before use in the predictive modelling to avoid detrimental effects of the skewness observed in the Cd concentration data (Bray et al., 2009). During model development, hyper-parametrisation of the different sensors tested was carried out using repeated k-fold cross-validation with 10 folds and 25 repeats, with the optimal set of parameters chosen based on the simplest model within the one standard error of the empirically optimal model (Breiman et al., 1984).

Prediction models were developed using PLS regression (Section 2.4.3.1; Table 2.11). The PLS regression used the R package *pls* (Mevik et al., 2020) wrapped in the *caret* package (Kuhn et al., 2021). The variable importance in projection (VIP; Section 2.4.4) scores from PLS model based on each sensor data were calculated using *varImp* function in the R package *caret* (Kuhn et al., 2021). Spectral regions including VIP with a score of more than 50% were identified as the main contributors for quantitative model development. The VIP scores are based on the weighted sum of the absolute regression coefficients for each variable scaled to have a maximum value of 100 (Kuhn et al., 2021).

4.2.5 Evaluation of model performance

The predictive accuracy of the different models was assessed and compared calculating RMSE, R^2, RPIQ (Bellon-Maurel et al., 2010), CCC (Lin, 1989), and bias. Details on performance statistics can be found in Section 2.4.5 of Chapter 2. These parameters were

calculated using R package *spectacles* (Roudier, 2021). The normalised RMSE value was calculated dividing RMSE by the mean Cd concentration (Pullanagari et al., 2016).

4.2.6 Other statistical analyses

Statistical summary of soil chemical properties (total Cd, total P, total C, total N, CEC, acid oxalate extractable Al, Fe, and Si, and pH) was calculated and tabulated into two groups: Allophanic and non-Allophanic soils. Data were checked for normality using the Shapiro-Wilk normality test. Skewed data were transformed (e.g., $\log_{10}$, square root) as needed to obtain normally distributed data. The effect of broad soil groups (Allophanic and non-Allophanic) on soil properties were analysed using a Welch's two-sample t-test. Correlation and regression analyses among relevant variables were performed, tabulated, and/or presented graphically as needed.

4.3 Results

4.3.1 Soil chemical properties

A summary of key soil chemical properties determined for the 50 Allophanic and 37 non-Allophanic soils used in this study is presented in Table 4.1. Total P and acid oxalate extractable Al concentrations in Allophanic soils (2.36 g P/kg and 3.09 g Al/kg soil, respectively) were higher ($p<0.001$) compared to non-Allophanic soils (1.30 g P/kg and 0.78 g Al/kg soil, respectively) (Table 4.1). Mean values of total C, total N, CEC, and soil pH were not significantly different ($p>0.05$) between the Allophanic and non-Allophanic soils (Table 4.1). Total soil C concentration ranged from 2.21 to 42.67 g/kg (Table 4.1). In the non-Allophanic soil group, Organic soils (27% of the total) showed higher total carbon content (>15 g/kg) reaching up to 42.6 g/kg (Table 4.1). Soil pH ranged from 4.81 to 6.44 (Table 4.1).

Table 4.1 Soil chemical properties of 50 Allophanic and 37 non-Allophanic topsoil (0–15 cm) samples collected from 30 New Zealand pastoral farms. For each soil property, results (p-values) from the Welch's two-sample t-test of both groups (Allophanic vs non-Allophanic) are included.

Chemical properties	Allophanic soil				non-Allophanic soil				p-value
	Min.	Mean	Max.	Skew.	Min.	Mean	Max.	Skew.	
Total Cd (mg/kg)	0.34	0.76	2.03	1.72	0.10	0.35	0.87	0.76	< 0.001
Total P (g/kg)	0.78	2.36	3.61	-0.17	0.46	1.30	2.49	0.26	< 0.001
Total C (%)	3.18	8.57	15.06	0.02	2.21	10.53	42.67	1.72	0.741
Total N (%)	0.34	0.83	1.31	-0.39	0.21	0.72	1.70	0.71	0.107
CEC (cmol/kg)[1]	15.03	33.40	48.21	-0.65	7.84	30.58	70.86	0.88	0.339
Al_o (g/kg)[2]	0.16	3.09	5.10	-0.49	0.18	0.78	1.48	0.07	< 0.001
Fe_o (g/kg)[3]	0.31	0.83	1.47	0.29	0.15	0.52	1.30	1.14	< 0.001
Si_o (g/kg)[4]	0.02	0.60	1.10	-0.31	0.01	0.08	0.41	2.74	< 0.001
Soil pH	4.95	5.61	6.08	-0.17	4.81	5.51	6.44	0.16	0.126

CEC (cmol/kg)[1] = Cation exchange capacity; Al_o (g/kg)[2] = Acid oxalate extractable Al; Fe_o (g/kg)[3] = Acid oxalate extractable Fe; Si_o (g/kg)[4] =Acid oxalate extractable Si.

4.3.2 Total Cd

The average total Cd concentration across all soils was 0.58 mg Cd/kg soil (Table 4.1). Total Cd concentration was significantly higher ($p<0.001$) in Allophanic soils (0.76 mg Cd/kg) compared to non-Allophanic soils (0.35 mg Cd/kg) (Table 4.1). Total Cd concentration in Allophanic soils was positively skewed (1.72) with 14% of samples containing more than 1.00 mg Cd/kg with a maximum concentration of 2.03 mg Cd/kg (Table 4.1). Total Cd concentration in non-Allophanic soils was also positively skewed (0.76) with 8% of non-Allophanic samples representing Organic (5%) and Pumice (3%) soil orders having a total Cd concentration of more than 0.60 mg Cd/kg reaching a maximum of up to 0.87 mg Cd/kg (Table 4.1).

A significant relationship ($R^2=0.58$, $p<0.001$) between total Cd and total P concentrations was found among all soil samples (Fig. 4.3). Total Cd concentration showed a significant positive correlation ($r\geq0.25$, $p<0.01$) with total C, total N, CEC, and acid oxalate extractable Al, Fe and Si, in soils (Table 4.2).

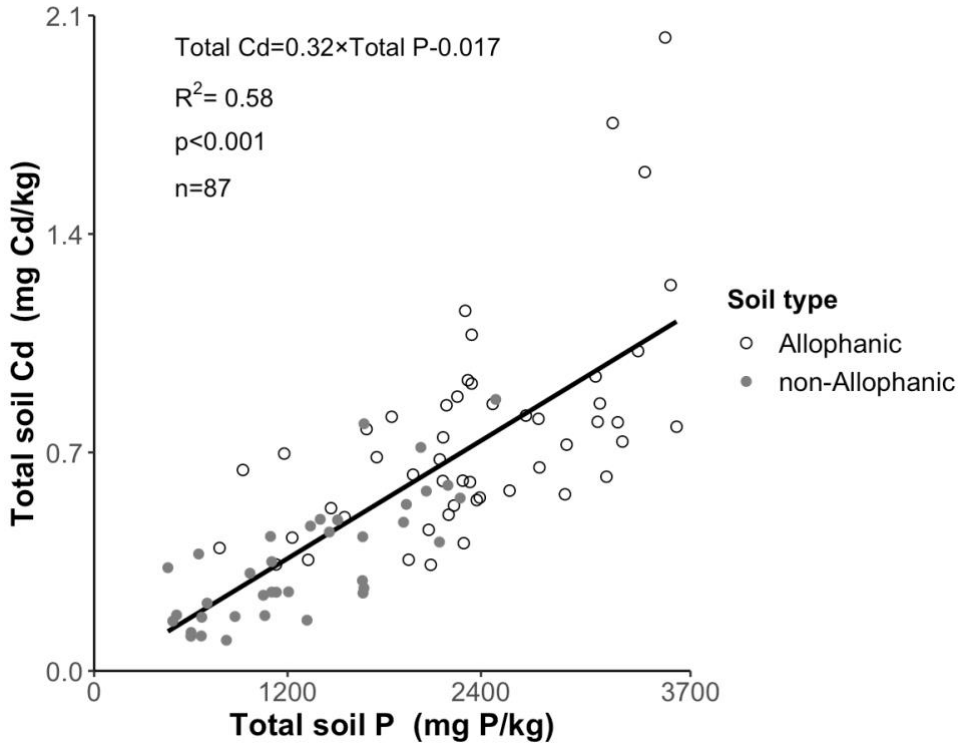


Fig. 4.3 Regression of total soil Cd in relation to total soil phosphorus (P) concentration for all topsoil samples i.e., Allophanic (white circles; n=50) and non-Allophanic (grey dots; n=37) soils from long-term pastural farms in New Zealand.

Total Cd showed a significant positive correlation with total P in both Allophanic (r=0.59, p<0.001) and non-Allophanic soils (r=0.78, p<0.001) and a significant positive correlation (r=0.42, p<0.01) with acid oxalate extractable Al in Allophanic soils (Table 4.2). However, no correlation was observed in non-Allophanic soils (Table 4.2). Total Cd also showed a significant positive correlation (r=0.44, p<0.001) with total C in non-Allophanic soils only (Table 4.2).

Table 4.2 Correlation analysis among total soil Cd and exchangeable, acid soluble, organic matter bound, metal oxides bound, and residual Cd with selected soil chemical properties measured for Allophanic (n=50) and non-Allophanic soils (n=37).

Chemical properties	Cadmium					
	Total	Exchangeable	Acid soluble	Metal oxides bound	Organic matter bound	Residual
All soils (n=87)						
P	0.80***	0.66***	0.79***	0.79***	0.67***	0.50***
C	0.42***	0.17	0.45***	0.29	0.57***	0.30**
N	0.52***	0.27*	0.53***	0.40**	0.63***	0.38***
CEC	0.43***	0.20	0.47***	0.30*	0.62***	0.35***
Al_o	0.53***	0.31**	0.51***	0.62***	0.43***	0.65***
Fe_o	0.29**	0.26*	0.24*	0.33**	0.04	0.45***
Si_o	0.50***	0.40***	0.49***	0.62***	0.30**	0.52***
pH	-0.07	-0.12	-0.15	0.00	-0.21	0.00
Allophanic soils (n=50)						
P	0.59***	0.55***	0.63***	0.67***	0.58***	0.60***
C	0.16	-0.04	0.28*	0.33	0.22	0.38
N	0.30*	0.04	0.34*	0.40**	0.26	0.48***
CEC	0.36**	0.09	0.43**	0.41**	0.35*	0.44**
Al_o	0.42**	0.21	0.48***	0.55**	0.43**	0.62***
Fe_o	0.13	0.05	0.17	0.14	-0.05	0.33*
Si_o	0.52***	0.26	0.52***	0.57***	0.46***	0.61***
pH	-0.24	-0.09	-0.28*	-0.26	-0.26	-0.11
non-Allophanic soils (n=37)						
P	0.78***	0.40*	0.81***	0.65***	0.70***	-0.01
C	0.44***	0.37*	0.75***	0.44**	0.76***	0.29
N	0.61***	0.40*	0.74***	0.40*	0.79***	0.26
CEC	0.49**	0.27	0.62***	0.28	0.75***	0.31
Al_o	0.07	-0.52**	-0.11	-0.02	0.10	0.53***
Fe_o	-0.20	-0.08	-0.25	-0.11	-0.28	0.35*
Si_o	-0.40*	-0.51**	-0.51**	-0.30	-0.45**	-0.05
pH	-0.19	-0.48**	-0.29	0.02	-0.33*	-0.02

Al_o=Acid oxalate extractable Al, Fe_o=Acid oxalate extractable Fe, Si_o=Acid oxalate extractable Si, CEC=Cation exchange capacity; p value <0.05=*, <0.01=**, and <0.001=***

4.3.3 Cadmium fractions

Sequential extraction separated Cd into one of five operationally defined fractions: exchangeable, acid soluble, metal oxides bound, organic matter bound, and residual Cd (Fig. 4.4). For all the soils studied, there was a significant relationship (R^2=0.90, p<0.001) between total soil Cd and sum of Cd extracted in the individual fractions (Fig. 4.4a). The average concentration of Cd in each of the fractions: exchangeable, acid soluble, organic matter bound, metal oxides bound, and residual Cd was 0.12, 0.11, 0.09, 0.12, and 0.02 mg Cd/kg soil, respectively (Fig. 4.4b).

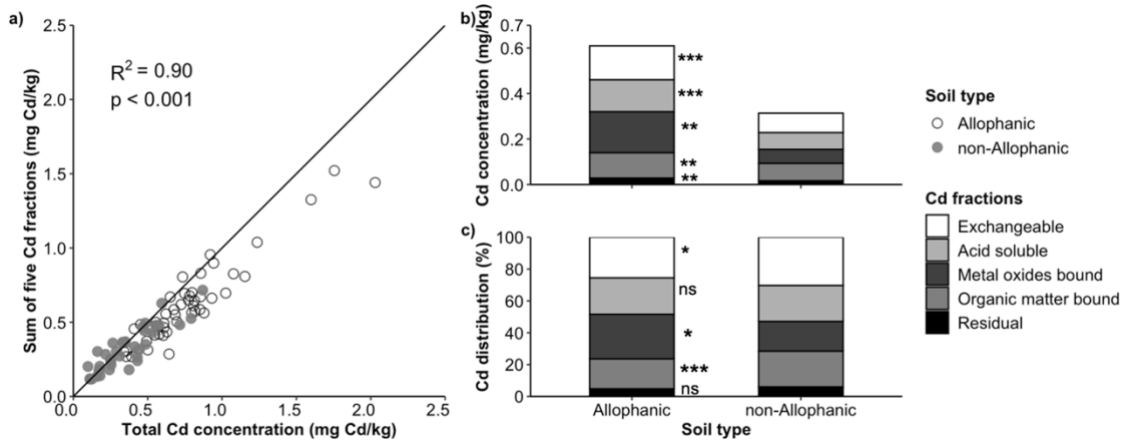


Fig. 4.4 Soil Cd sequential extraction analysis: a) total soil Cd against the sum of five Cd fractions for Allophanic (white circles) and non-Allophanic (grey dots) soils. The diagonal centre line from the origin is a 1:1 line. b) the mean distribution of Cd between the fractions expressed as concentration (mg Cd/kg), and c) the distribution of Cd between the fractions expressed as a percentage of the sum of five Cd fractions for Allophanic and non-Allophanic soils (n=87). Significance of mean comparison is presented as $p<0.05$=*, $p<0.01$=**, and $p<0.001$=***.

On average, Cd concentration in all the fractions were significantly higher ($p<0.01$) in Allophanic soils than non-Allophanic soils (Fig. 4.4b). The metal oxides bound Cd fraction average concentration was higher (0.18 mg Cd/kg soil) than other fractions in Allophanic soils (Fig. 4.4b). Exchangeable Cd average concentration (0.09 mg Cd/kg soil) were higher than other fractions in non-Allophanic soils (Fig. 4.4b). On average, both Allophanic and non-Allophanic soils showed very low amounts of average residual Cd (0.03 mg Cd/kg soil and 0.02 mg Cd/kg soil, respectively) (Fig. 4.4b).

For all soils studied, a substantial proportion of Cd (71.3%) was found in the exchangeable and potentially-available forms i.e., acid soluble and organic matter bound, while about 23.3% was bound to metal oxides (Fig. 4.4c). The proportion of Cd in the residual fraction was always very low (5.4%; Fig 4.4c). Comparing Cd fractions between the Allophanic and non-Allophanic soils, metal oxides bound Cd was significantly higher ($p<0.001$) in Allophanic soils (27.9%) than non-Allophanic soils (18.6%) (Fig. 4.4c). Exchangeable and organic matter bound Cd were significantly higher ($p<0.05$) in non-

Allophanic soils (30.4% and 22.5%, respectively) compared to Allophanic soils (25.6% and 18.9%, respectively) (Fig. 4.4c).

For all soils combined, the Cd concentration in all five tested fractions showed a significant ($p<0.05$) positive correlation with total P ($r\geq0.50$), acid oxalate extractable Al ($r>0.30$), Si ($r>0.30$), and total N ($r>0.25$) (Table 4.2). Acid soluble, organic matter bound, metal oxides bound, and residual Cd showed a significant positive correlation (($r\geq0.30$, $p<0.05$) with the CEC of soils (Table 4.2). Acid soluble, metal oxides bound, and residual Cd showed a significant positive correlation ($r>0.35$, $p<0.001$) with total N (Table 4.2). Acid soluble, metal oxides bound, and residual Cd showed a significant positive correlation ($r>0.20$, $p<0.05$) with acid oxalate extractable Fe (Table 4.2). Acid soluble, organic matter bound, and residual Cd showed a significant positive correlation ($r\geq0.30$, $p<0.05$) with total C (Table 4.2).

4.3.4 Spectral similarity

The principal component analysis of vis-NIR, MIR, and pXRF spectra for the first two principal components showed the spread of Allophanic and non-Allophanic samples as per spectra characteristics (Fig. 4.5). The PCA of vis-NIR spectra showed the greater separation between Allophanic and non-Allophanic soils, while in the MIR spectra allophanic soil showed low spectral variability and they were almost similar to non-Allophanic soil (Fig. 4.5). The first two principal components of pXRF spectra explained 90% spectral variance than MIR (74%) and vis-NIR (63%) (Fig. 4.5).

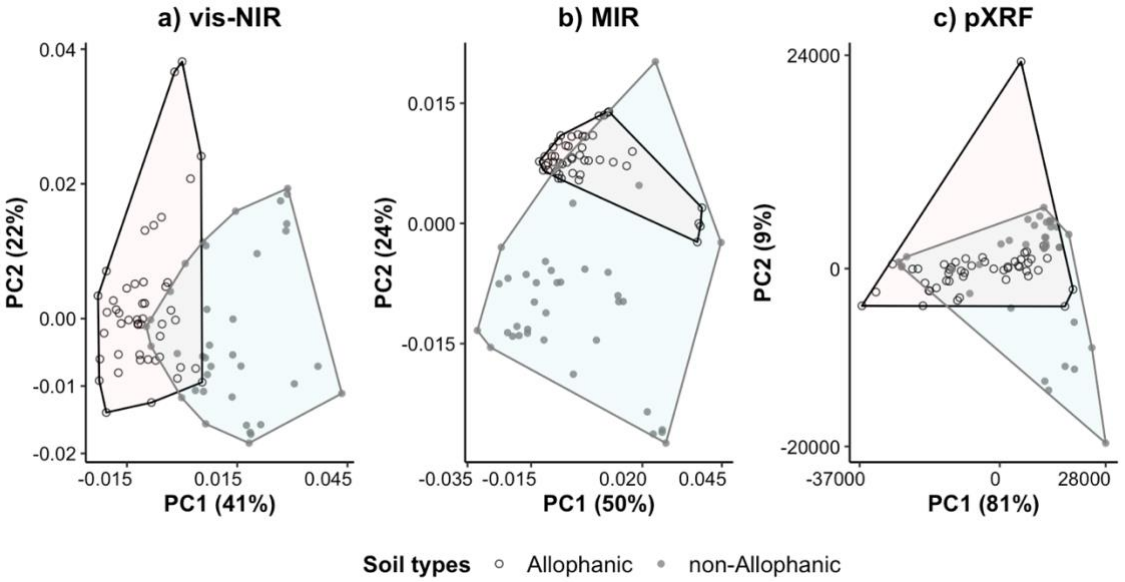


Fig. 4.5 Principal component analysis (PC1 and PC2) of a) vis-NIR, b) MIR, and c) pXRF for Allophanic (n=47, white circles) and non-Allophanic (n=37, grey dots) soils. Spectral variance explained by each principal component is shown in % value inside the bracket.

4.3.5 Model predictions of total Cd and Cd fractions

To quantify total Cd and Cd present in each of the five fractions considered, prediction models were developed using proximal sensor data from vis-NIR, MIR, and pXRF sensors independently as input for the PLS algorithm to quantify total Cd and Cd fractions (Figs 4.6 and 4.7; Table 4.3). In general, PLS models vis-NIR data as input outperformed those models based on MIR and pXRF data to quantify total, acid soluble, and organic matter bound Cd (Figs 4.6 and 4.7; Table 4.3). This is evidenced, after cross-validation, by reduced $RMSE_{cv}$, $nRMSE_{cv}$, and $bias_{cv}$, and increased $R2_{cv}$, CCC_{cv}, and $RPIQ_{cv}$, suggesting improved acccuracy of estimations of total Cd and Cd fractions (Figs 4.6, 4.6; Table 4.3). Models using MIR or vis-NIR data outperformed pXRF data to predict exchangeable Cd (Figs 4.6, 4.7b, and 4.7e; Table 4.3). The performance of PLS models, using proximal sensors considered in this study, was relatively poor when quantifying both metal oxides bound and residual Cd (Figs 4.6, 4.7d, and 4.7f).

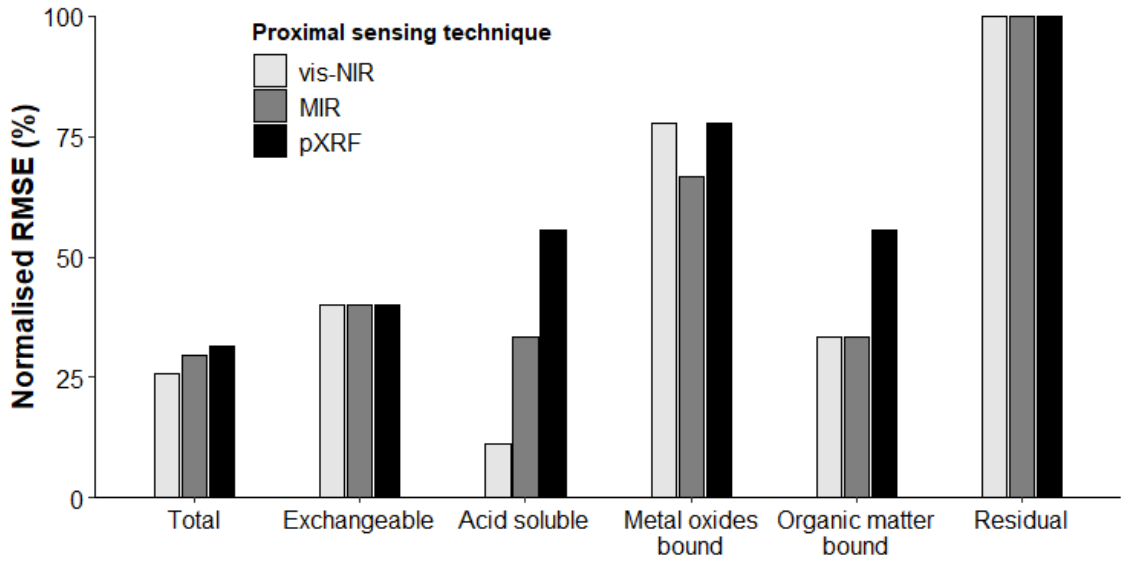


Fig. 4.6 Normalised root mean square error (nRMSE) percentage for model prediction of total soil Cd and five Cd fractions: exchangeable, acid soluble, metal oxides bound, organic matter bound, and residual Cd using individual pXRF, vis-NIR, MIR proximal sensor data as input for PLS.

Prediction model using vis-NIR data as input quantified total Cd with $nRMSE_{cv}$ of 26% and CCC_{cv} of 0.85 (Figs 4.6 and 4.7a; Table 4.3). Models using either MIR or pXRF data also predicted total Cd with less success than the model using vis-NIR data ($nRMSE_{cv}$ of 30% and 31% and CCC_{cv} of 0.77 and 0.75, respectively; Figs 4.6 and 4.7a; Table 4.3). Models using MIR or vis-NIR data assessed exchangeable Cd with $nRMSE_{cv}$ of 40% and CCC_{cv} of 0.57 (Figs 4.6b and 4.7b; Table 4.3). For acid soluble and organic matter bound Cd, the best prediction results were obtained by models using vis-NIR data (Figs 4.6, 4.7; Table 4.3). These optimal models for acid soluble and organic matter bound Cd performed relatively well, with $nRMSE_{cv}$ of 11% and 33% and CCC_{cv} 0.97 and 0.84, respectively (Figs 4.6, 4.7c, and 4.7e; Table 4.3).

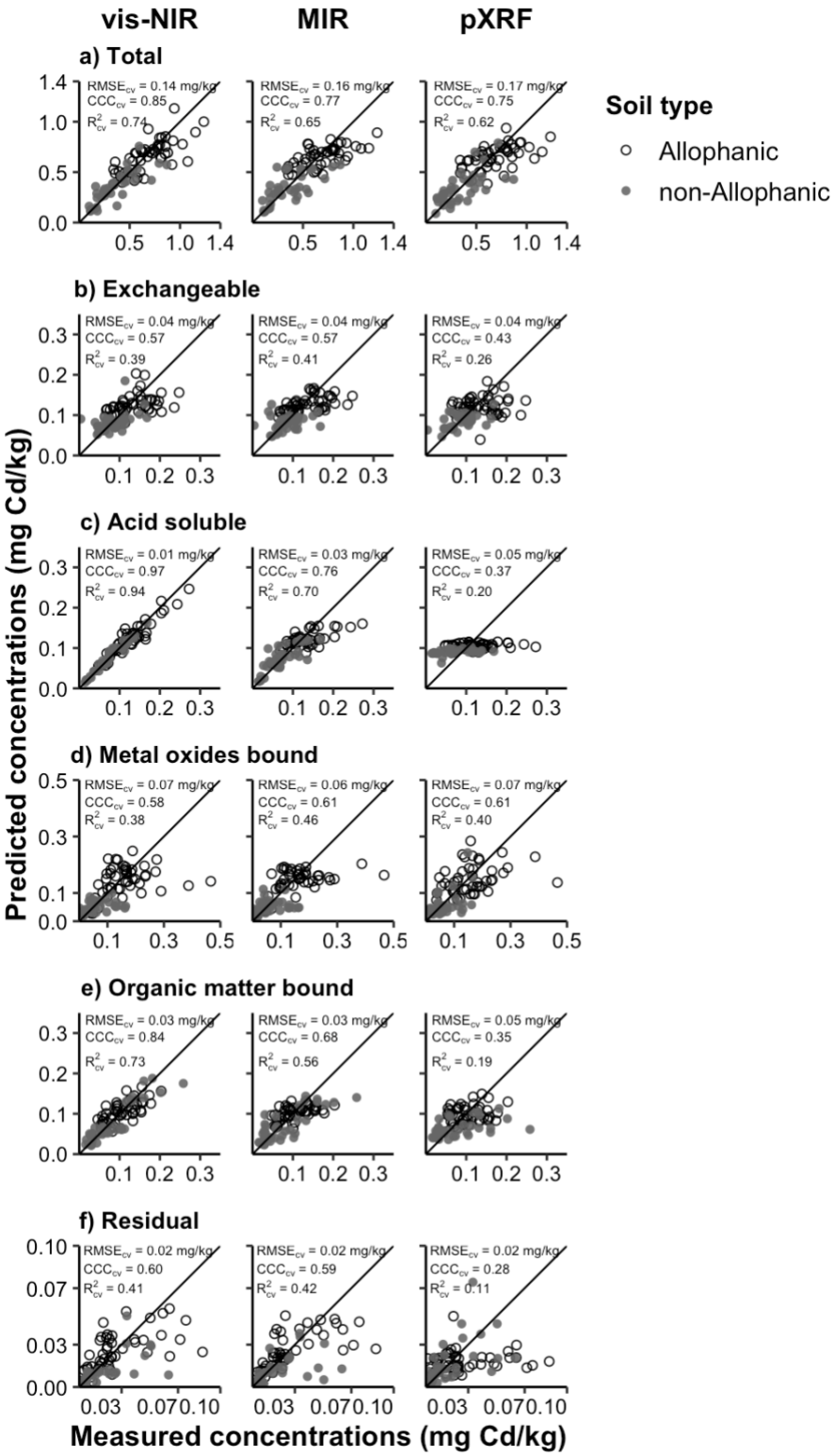


Fig. 4.7 Measured vs predicted concentrations of a) total, and Cd fractions in b) exchangeable, c) acid soluble, d) metal oxides bound, e) organic matter bound, and f) residual Cd for the cross-validation (cv) set (n=84) including Allophanic (white circles) and non-Allophanic (grey dots) soils by models using vis-NIR, MIR, and pXRF spectral data as input for PLS. A black diagonal line going through the origin is a 1:1 line.

Table 4.3 Cross-validation results of models predicting total soil Cd and Cd fractions in exchangeable, acid soluble, metal oxides bound, organic matter bound, and residual Cd using vis-NIR, MIR, and pXRF spectral data as input for PLS.

Cd	vis-NIR					MIR					pXRF				
	RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias
Total	0.14	0.74	2.94	0.85	-0.017	0.16	0.65	2.51	0.77	-0.025	0.17	0.62	2.41	0.75	-0.026
Exchangeable	0.04	0.39	1.90	0.57	-0.007	0.04	0.41	1.95	0.57	-0.007	0.04	0.26	1.73	0.43	-0.008
Acid soluble	0.01	0.94	5.20	0.97	<0.001	0.03	0.70	2.33	0.76	-0.004	0.05	0.20	1.42	0.37	-0.01
Metal oxides bound	0.07	0.38	1.69	0.58	-0.014	0.06	0.46	1.81	0.61	-0.01	0.07	0.40	1.72	0.61	-0.011
Organic matter bound	0.03	0.73	2.81	0.84	-0.004	0.03	0.56	2.15	0.68	-0.006	0.05	0.19	1.58	0.35	-0.01
Residual	0.02	0.41	1.27	0.60	-0.003	0.02	0.42	1.28	0.59	-0.003	0.02	0.11	0.99	0.28	-0.005

Performance statistics of the optimal models are in bold letter.

Variable importance in projection highlighted the importance of specific regions of vis-NIR, MIR, and pXRF spectra for quantitative predictions of total Cd and Cd in each of the fractions considered (Fig. 4.8). For predicting total Cd, relevant variable importance in projection included: (1) in the vis-NIR spectra: the 400–850 nm region associated with soil organic matter and Fe oxides and the 2200–2500 nm region associated with Al– and Fe–containing minerals and soil organic matter (Fig. 4.8a), (2) in the MIR spectra: the 3700–3000 cm^{-1} region associated with Al– and Fe–containing minerals and the 2130–1700 cm^{-1} region that is due to metal–carbonyl interactions (Fig. 4.8a), and (2) considering the pXRF spectra; the regions at 3.5–4.2 keV and 13.5–14.5 keV were of relevance (Fig. 4.8a).

For predicting exchangeable Cd, the VIP highlighted the 3700–3000 cm^{-1} (Al– and Fe–containing minerals) and 2130–1700 cm^{-1} (metal–carbonyl) regions of the MIR spectra as important regions (Fig. 4.8b). For predicting acid soluble Cd, the 2200–2500 nm (Al– and Fe–containing minerals and soil organic matter complex) regions and 400–850 nm (soil organic matter and Fe oxides) regions of the vis-NIR spectra were important (Fig. 4.8c). For predicting organic matter bound Cd, the 400–850 nm (soil organic matter and Fe oxides) and 2200–2500 nm (Al– and Fe–containing minerals and soil organic matter) regions of the vis-NIR spectra were important (Fig. 4.8e).

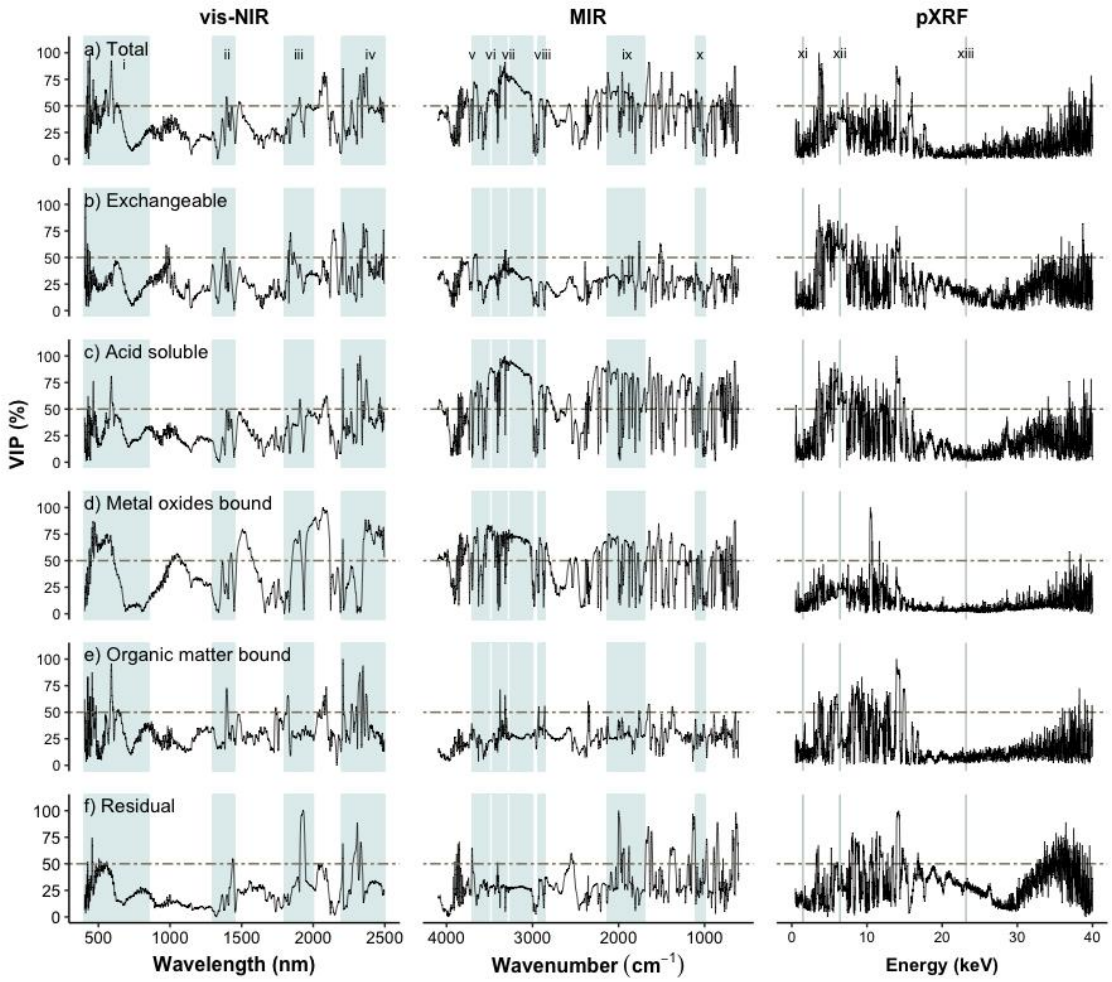


Fig. 4.8 Variable importance in projection (VIP) highlighting the importance of the selected regions in the vis-NIR, MIR, and pXRF spectra to quantify a) total Cd and Cd fractions in b) exchangeable, c) acid soluble, d) metal oxides bound, e) organic matter bound, and f) residual Cd. The shaded area in the vis-NIR region: i) soil colour, iron oxides, and soil organic matter (400−850 nm), ii) minerals (1300−1450 nm), iii) Al– and Fe–containing minerals and soil organic matter (1800−2000 nm), and iv) Al– and Fe–containing minerals and soil organic matter complex (2200−2500 nm); in the MIR region: v) Al– and Fe–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) and x) water (3484 and 3278 cm^{-1}); and in the XRF region: xi) Al (1.48 keV), xii) Fe (6.40 keV), and xiii) Cd (23.17 keV).

4.4 Discussion

Total soil Cd concentration reported in this study agreed with the total Cd concentration range found in similar New Zealand topsoils (Table 4.1; Abraham (2018); Salmanzadeh et al. (2016); Stafford (2017)). A significant positive relationship between total soil Cd and total soil phosphorus concentrations was found which is again in agreement with other studies in New Zealand agricultural soils suggesting phosphate fertiliser application as the primary source of Cd accumulation above background concentration (Fig. 4.3; Table 4.2; Longhurst et al. (2004); Zanders et al. (1999)).

The pedogenic origin of the soil samples also influenced the total soil Cd concentration (Fig. 4.3; Gray et al. (1998); Stafford et al. (2018c)). Allophanic and Pumice soils (a non-Allophanic soil order) formed from volcanic parent materials showed more Cd than other soil orders (Table 4.1; Appendix 5.1; Cavanagh (2014); Hewitt et al. (2021)). At the same time, Organic soils (a non-Allophanic soil order) rich in soil organic matter had higher total Cd compared to other non-Allophanic soils (Table 4.1; Appendix 5.1; Taylor et al. (2007)). In addition to total soil Cd concentration, soil pedogenesis should be considered for the successful implementation of cadmium management strategy such as the TFMS in New Zealand (Stafford et al., 2018c).

4.4.1 Factors affecting total soil Cd and Cd fractions

For the soil samples used in this study, which were from pastoral farms with long-term phosphate fertiliser application history, the main soil properties affecting Cd accumulation and distribution were Al containing minerals and soil organic matter (Table 4.2; Loganathan et al. (2012); Yuan and Lavkulich (1997)). In Allophanic soils, Cd can be found sorbed to poorly ordered alumino-silicate clay minerals (such as allophane and imogolite) which contain high variable–charge (pH-dependent) and high specific sorption capacity (Table 4.2; Christensen and Haung (1999); Reiser et al. (2014)). This may be due to than non-Allophanic soils for pasture production (Hewitt, 2010). This characteristic of Allophanic soils contributes to the higher total P and Cd concentrations in Allophanic soil than non-Allophanic soil (Table 4.1). Cadmium can also be held on terminal functional groups of variable charged soil organic matter by electrostatic attraction or covalent bonding (Bolan et al., 2003b; Young, 2013). Cation exchange capacity is therefore known to influence the mobility of Cd sorbed to soil organic matter

(Table 4.2; Amacher et al. (1986); Stafford et al. (2018c)), and despite soil pH being a major determining factor of CEC, soil pH had no significant influence in total soil Cd concentration within the topsoils (Table 4.2; Salmanzadeh et al. (2016)).

Though total Cd concentration was elevated in Allophanic soils, the proportion of Cd distributed between the plant-available and potentially-available fractions (i.e., exchangeable, acid soluble, and organic matter bound Cd) was relatively low when compared to that distribution of Cd fractions in non-Allophanic soils (Figs 4.4b and 4.4c). This effect of soil type on Cd availability is not considered yet in the New Zealand TFMS (Fertiliser Association, 2019; Stafford et al., 2018c). Previous studies showed that application of an equal amount of phosphate fertiliser to Allophanic and non-Allophanic soils resulted into more Cd availability in the non-Allophanic than Allophanic soils (Stafford et al., 2018a; Yi et al., 2020). This could lead to higher Cd accumulation in plant edible parts grown in non-Allophanic soils compared to Allophanic soils causing more risk of food chain transfer (Chapter 6; Cavanagh et al. (2019); Gray et al. (2019a)).

This study, for the first time, has studied Cd fractions in a large number of topsoil representative of soil orders found in New Zealand agricultural soils (n=87) representing multiple soil orders (Appendix 4.1) compared to much smaller sample sizes (n≤30) in the previous studies in New Zealand (Fig. 4.4; Table 2.5; Gray et al. (2000); Salmanzadeh et al. (2016)). The distribution of Cd fractions in this set of soils from different parts of New Zealand were similar to the findings by Salmanzadeh et al. (2016) using only two soil orders (Allophanic and Gley) in the Waikato region (Table 2.5; Fig. 4.4).

Exchangeable Cd concentration recovered were high in range in the present study than in the previous studies recovering a wide range of 3–34% (Fig. 4.4c; Table 2.5; Bolan et al. (2003a); Gray et al. (2000)) and this may be due to weathering of Al– and Fe–containing minerals and decomposition of organic matter bound Cd as a function of acidic soil pH, long-term exposure in the field environment and continuous addition of phosphate fertiliser for crop production (McBride and Li, 2022; Oliver et al., 2006). A poor correlation was observed between exchangeable Cd and acid oxalate extractable Al, Fe, and Si, total C and total N, respectively (Table 4.2; Kim and Fergusson (1991)).

In Allophanic soils, the most significant proportion of total soil Cd was associated with the metal oxides bound fraction (Fig. 4.4c) and this may be a function of poorly ordered alumino-silicate clay minerals abundant in Allophanic soils showing high specific

sorption capacity to retain Cd (Fig. 4.4c; Table 4.2; Bolan et al. (2003a); Mann and Rate (1998)). This was indicated by significant correlations of metal oxides bound Cd with acid oxalate extractable Al, Fe, and Si (Table 4.2; Krishnamurti et al. (1995); Salmanzadeh et al. (2016)).

The residual Cd concentration in this set of soils was very low (average 0.02 mg Cd/kg soil; Fig. 4.4b), particularly in comparison to the residual Cd concentration detected in other studies (0.05–0.4 mg Cd/kg soil) (Fig. 4.4c; Table 2.5; Gray et al. (1999b); Gray et al. (2000)). Accumulation of Cd in the soil residual fraction in recent years may have decreased due to a shift from high to low Cd sources of phosphate fertiliser since the 1990s (Abraham, 2018). The data in this Chapter show that correlation coefficient value between total P and residual Cd was lower (r=0.50) than other fractions (r>0.65) (Table 4.2; Kashem et al. (2011)).

4.4.2 Predictions of total Cd in agricultural soils using proximal sensing

The present study quantified total Cd concentration in agricultural soils using the three proximal sensing techniques: vis-NIR, MIR, and pXRF sensor. Rapid quantification of total Cd using proximal sensing techniques requiring minimum sample preparation has the potential for cost-effective analysis of total Cd at individual farms to successfully implement Cd monitoring programmes such as the TFMS in New Zealand (Fertiliser Association, 2019). Among the three proximal sensing techniques, the vis-NIR sensor is the most cost-effective instrument and has the highest accuracy (Fig.4.7a; Table 4.3; Nawar et al. (2019)). With inexpensive analysis cost for soil Cd compared to reference laboratory methods, more samples can be analysed with large spatial coverage for less expense (Shepherd and Walsh, 2007). This can be useful in generating spatial associations of soil Cd distribution across farms (Stafford et al., 2018c). Using proximal sensors in this way, a farm spatial aspect could be added to Cd management for more effective implementation of the New Zealand TFMS (Gray and Cavanagh, 2022).

This study directly quantified total Cd using vis-NIR spectra as the input for PLS algorithm (Figs 4.6a and 4.7a; Table 4.3). A previous study by (Stafford et al., 2018b) performed indirect quantification of Cd based on the significant relationship between Cd and vis-NIR predicted soil total C and total N. The quantification of total Cd based on

vis-NIR relied on the association of Cd with spectrally active soil components: (1) soil organic matter, and (2) Al– and Fe–containing minerals (Kooistra et al., 2001; Song et al., 2012; Xu et al., 2020). In the current study, this association is indicated by (1) variable importance in projection of selected regions including spectral response from organic matter and minerals (Fig. 4.8a); and (2) a significant correlation between total Cd and total C, total N, acid oxalate extractable Al and Fe (Table 4.2; Christensen and Haung (1999); Gholizadeh et al. (2021); Young (2013)). The Cd quantification based on MIR relied on the association of Cd with Al– and Fe–containing minerals and metal–carbonyl bonds in the matrix (Janik et al., 1998; Wang et al., 2017), and this association is supported by variable importance in projection of the selected regions including spectral response from Al– and Fe–containing minerals and metal–carbonyl bonds (Fig. 4.8a). For pXRF, soil Cd concentration was below the built-in calibration model lower detection limit of 5 mg Cd/kg soil (Table 4.1). Therefore, the Cd prediction model was developed using pXRF spectra as input, which directly correlates with Cd concentration in the matrix (Chapter 3).

4.4.3 Assessing the distribution of soil Cd fractions using proximal sensing

This study tested, for the first time, whether proximal sensors using individual vis-NIR, MIR, and pXRF, predict Cd fractions in agricultural soils of New Zealand (Figs 4.6 and 4.7; Table 4.3). The reflectance spectroscopy techniques showed the feasibility of predicting specific Cd fractions in agricultural soils. A model using vis-NIR data showed optimal performance to quantify Cd in the acid soluble and organic matter bound fractions (Figs 4.6 and 4.7; Table 4.3). A model using MIR data predicted exchangeable Cd relatively well (Figs 4.6 and 4.7; Table 4.3). In general, the metal oxides bound and residual Cd were predicted poorly or very poorly (Fig. 4.6; Table 4.3). Soil organic matter might have caused noise in the pXRF spectra with the least effect observed on the metal oxides bound fraction (Table 4.2; Fig. 4.8; Ravansari et al. (2020)).

Accurate prediction of the acid soluble and organic matter bound Cd is relevant to the study of plant Cd accumulation as both are potentially-available fractions for plant uptake (Chavez et al., 2016; El-Mufleh Al Husseini et al., 2013; Gray et al., 2000). The accuracy achieved for exchangeable Cd in this study can be feasible to classify agricultural soils

with respect to the tier-based systems (Fig. 4.7b; Cécillon et al. (2009)). These results show the potential of using proximal sensing techniques for monitoring soil Cd fractions to minimise plant Cd uptake (Cipullo et al., 2019).

The association of acid soluble and organic matter bound Cd with soil organic matter contributed to the prediction of these fractions using vis-NIR spectra (Figs 4.6, 4.8c, and 4.8e; Tables 4.2 and 4.3; Gray et al. (2000); Kooistra et al. (2001); Krishnan et al. (1980)). The average accuracy of exchangeable Cd prediction in the present study (Figs 4.6 and 4.7b; Table 4.3) agrees with previous studies that have quantified exchangeable TE including As, Cu, and Zn in contaminated soils (Chakraborty et al., 2017; Chodak et al., 2007). An association between exchangeable Cd and Al– and Fe–containing minerals and/or soil organic matter may explain the prediction accuracy of Cd in this fraction (Fig. 4.8b; Stenberg and Rossel (2010); Wang et al. (2017)). In the present study, exchangeable Cd was poorly correlated to acid oxalate extractable Al and Fe and total C and total N (Fig. 4.7b; Table 4.2; Kim and Fergusson (1991)). Despite air-drying, the variable importance in projection (VIP) highlighted a masking effect by water at ~3484 and ~3278 cm^{-1} of the MIR spectra (Fig. 4.8b; Appendix 2.3). This may have been due to a water bridging mechanism which could explain the poor accuracy of exchangeable Cd prediction (Fig.4.8b; Chakraborty et al. (2017); Sposito (1984)). In addition, the presence of exchangeable Cd in inorganic forms (e.g., $CdCl^{+}$, $Cd(NO_3)_2$), which have no influence on reflectance spectra, might have reduced the accuracy of using reflectance spectra (Bolan et al., 1999; Li et al., 2021; Moros et al., 2009).

Metal oxides bound and residual Cd fractions were not accurately predicted using the proximal sensing techniques deployed (Figs 4.6e and 4.6f; Table 4.3; Chakraborty et al. (2017)). The very low concentration range of the residual Cd fractions limited the possibility of detecting and predicting this TE fraction using a proximal sensor (Figs 4.7d and 4.7f; Xie et al. (2012)). Residual Cd reflects both Cd from parent materials and that historically has moved from other fractions into the residual fraction with time (Abraham, 2018; Kubier et al., 2019).

4.5 Conclusions

Rapid quantification of agricultural total soil Cd concentration (<2 mg Cd/kg soil) using either of the three proximal sensing techniques including vis-NIR, MIR, and pXRF

spectroscopy with minimum sample preparation could potentially be deployed for cost-effective analysis of total Cd from pastoral farm samples for effective implementation of Cd monitoring programmes such as the TFMS in New Zealand. Reflectance spectra captured the association of Cd with spectrally active soil components for accurate assessment of total Cd and distribution of Cd in available and potentially-available fractions (i.e., exchangeable, acid soluble, and organic matter bound Cd). Reflectance spectroscopy techniques showed potential to monitor the distribution of Cd fractions in agricultural soil to minimise plant Cd uptake and subsequent food chain transfer risks. Soil spectral libraries developed at a regional or national scale can potentially reduce the number of farm samples required to be sent for reference laboratory analysis, thereby underpinning the wider implementation of proximal sensing techniques.

Chapter 5

Rapid analysis of local soil cadmium concentration using a regional soil spectral library

Abstract

Proximal sensors including vis-NIR, MIR, and pXRF sensors have been proposed as rapid and cost-efficient techniques to monitor total soil Cd concentration. Proximal sensor-based Cd analysis can benefit from soil spectral libraries (SSLs) covering large extensions (e.g., regional, as in the Chapter 3). However, the prediction accuracy of these large libraries is of limited use when directly applied to targeted local scales such as agricultural soils (e.g., farm-scale, see Chapter 4). In this Chapter, the regional SSL (from soils covering Otago and Southland regions) was used to assess the Cd concentration in a local set (from agricultural soils taken more locally). Multivariate statistical analysis including PLS regression and LOCAL algorithms were used to find the optimal strategy that combines spectral information and total Cd concentration data. Prediction models based on the regional SSL and subsets selected either by (1) spectral and/or (2) land use similarity with the local set and also spiking with selected local set samples, were developed to assess the Cd concentration in the local set. The prediction model using MIR data of the regional SSL pastoral soil subset (n=283) spiked with 12 local samples weighted (×4) as input for LOCAL algorithm performed optimal to quantify local Cd concentration with an RMSE of 0.22 mg Cd/kg, CCC of 0.78, and RPIQ of 1.93. This study developed a systematic strategy to customise models based on proximal sensors to accurately quantify Cd at a farm scale while leveraging a SSL containing data from three proximal sensors. The large SSL can be used to quantify total soil Cd concentration in the TFMS with relatively little extra effort as implemented in New Zealand. The gap of soil orders and Cd concentration range between the SSL and local set should be filled by spiking SSL with selected local samples for accurate local Cd prediction.

5.1 Introduction

Cadmium is a soil TE, which at elevated plant-available concentration can transfer through the food chain potentially causing health, environmental, and economic risks (Alloway, 2013). The historical application of phosphate fertiliser containing trace amounts of Cd to New Zealand agricultural soils has underpinned the implementation of a TFMS to monitor and manage Cd accumulation in agricultural soils (Abraham, 2018; CWG, 2011). Proximal sensing techniques, including vis-NIR, MIR, and pXRF spectroscopy, either independently or combined via GRA, predicted soil Cd concentration (Chapters 3 and 4); thus there is an opportunity to reduce the cost of analysis when it is of relevance monitoring Cd at contrasted spatial scales (Nawar et al., 2019; Shrestha et al., 2022).

Visible-NIR and MIR sensors have been used to quantify multiple soil properties, including soil C and N, at a local scale (Nawar and Mouazen, 2017b; Sanderman et al., 2021; Sankey et al., 2008). The reflectance spectra have been used to quantify soil Cd depending on Cd co-varying with spectrally active soil components including soil organic matter and Al– and Fe–containing minerals (Soriano-Disla et al., 2014). In contrast, pXRF sensor quantifies soil Cd (and other elements) using a built-in calibration model depending on the concentration (Padilla et al., 2019) given that it is above the detection limit of the instrument used (e.g., 5 mg Cd/kg soil for the Olympus Vanta C series instrument; Table 2.6) (Rouillon and Taylor, 2016). The choice of proximal sensing technique used is determined by cost, portability, and the precision required (Nawar et al., 2019).

Besides, long-term repetitive Cd analysis of samples in an accredited reference laboratory required under the TFMS necessitates time-consuming destructive sample processing using strong acids (Fertiliser Association, 2019; Nocita et al., 2015). Soil analysis using proximal sensors is non-destructive, rapid, and requires minimum sample preparation without any chemical treatments and is hence cost-effective (Nduwamungu et al., 2009). The associated savings could be used to increase sampling intensity over space and time to include spatial soil heterogeneity (Nocita et al., 2015).

The development of SSL is the basis of the wider application of proximal sensing techniques to environmental monitoring (Shepherd and Walsh, 2002). Soil spectral libraries are obtained by compiling spectral, physical, chemical, biological, and spatial

information on soil samples to understand the relationship between soil properties and their spectral response (Viscarra Rossel et al., 2016). Large-scale SSLs can be used to develop calibration models for local applications reducing the number of local samples to be analysed by reference laboratory method (Brown, 2007; Sila et al., 2016).

Large SSLs, including regional/national, continental, and global coverage, show limitations in their ability to accurately quantify soil properties at very local scales (Gogé et al., 2014; Kuang and Mouazen, 2013). The SSL may be limited by (1) spectral characteristics, (2) land use, (3) soil types, and (4) the range of values for soil property in concern compared to the local set. There are different strategies to overcome such limitations while developing a calibration model based on the large SSL to assess soil property in a local set: (1) targeted selection of a subset of samples from the large SSL; (2) spiking the large SSL using selected samples from a local set (with or without weights); and (3) selection of algorithm (Brown et al., 2006; Ramirez-Lopez et al., 2013; Wetterlind and Stenberg, 2010).

Different methods for the selection of the best subset of samples from the large SSL have been proposed based on: (1) spectral similarity, (2) stratified sampling, and (3) land-use type (Lobsey et al., 2017; Shi et al., 2015). Gomez et al. (2020) used a SSL subset with similar MIR spectra to target samples to predict soil organic carbon, achieving a reduction of RMSE (from 16 to 4.5 g C/kg soil) with an increased R^2 (from 0.64 to 0.97). Moura-Bueno et al. (2020) stratified vis-NIR SSL spectra by land-use type to quantify soil organic carbon content in local samples with an improved accuracy of RMSE (reduced from 1.02% to 0.67%) and R^2 (increased from 0.76 to 0.88).

Spiking, on the other hand, is a procedure that adds a few representative samples from local set samples into the large SSL filling the gap of spectral characteristics, land use, soil type, and/or concentration range (Sankey et al., 2008). When spiking a large SSL, a proportion of local samples are still analysed using reference laboratory methods, which incurs extra cost (Hong et al., 2018). Two approaches have been proposed to add local samples into SSL-based models: (1) select spectrally similar local samples by principal component analysis (PCA) and (2) use of concentration values to select local samples to fill the gap (Guerrero et al., 2010). The optimal number of local samples required to develop augmented SSL-based models can range from 15% to 70% for regional to continental level SSL, respectively (Li et al., 2020; Nawar and Mouazen, 2017a). Finally,

besides spiking, giving weights (i.e., multiple entries of the same sample) to the added local samples can also be a cost-effective solution to increase the representativeness of local samples in the SSL-based models (Greenberg et al., 2022).

Algorithms have also shown influence in predicting local sample characteristics using large SSL (Brown, 2007; Gogé et al., 2014; Ramirez-Lopez et al., 2013). Partial least squares regression is the most frequently used algorithm (Li et al., 2020; Nocita et al., 2014). Memory-based learning algorithms, such as LOCAL, have also been gaining popularity due to their ease of implementation, computational speed, and accuracy (Sanderman et al., 2021; Shenk et al., 1997; Summerauer et al., 2021).

Cadmium monitoring in agricultural farms can benefit from developing SSL covering large areas (e.g., regional, as in Chapter 3). However, the predictive performance of these SSLs is poor when directly applied to targeted agricultural soils (e.g., farm-scale, see Chapter 4). This Chapter addresses: 1) how to use SSL for predicting Cd concentration in local samples by applying spectral and land use similarity and 2) how to spike SSL to fill the gap of representativeness of local samples to quantify Cd concentration in local samples by vis-NIR, MIR, and pXRF spectroscopy. This chapter used the regional SSL from the Chapter 3 and the local set of agricultural soils from the Chapter 4 to achieve these objectives.

5.2 Materials and methods

5.2.1 Regional soil spectral library and local set

The regional SSL for Otago and Southland (Chapter 3) contains topsoil (0−20 cm) samples (n=625) collected at a regular 8 km spacing covering c. 40,000 km^2 between sea level and 2,000 m above sea level within the Otago and Southland regions, New Zealand (Appendix 5.1; Martin et al. (2016); Shrestha et al. (2022)). Samples within this SSL were collected from pasture (45%) and other land uses (55%) (Appendix 5.1). The dominant soil types across the SSL were non-Allophanic soils including Brown (53%) and Pallic (24%) soils (Fig. 5.1a; Appendix 5.1; Hewitt et al. (2021)). The total soil Cd concentration range across samples in this SSL was 0.005−1.31 mg Cd/kg with an average of 0.08 mg Cd/kg (Fig. 5.1b; Appendix 5.1).

The local set (Chapter 4) contained topsoil (0−15 cm) samples (n=87) collected from 30 commercial pastoral farms with long-term phosphate fertiliser application history across the North Island (22) and South Island (8) in New Zealand. Details on soil sampling sites and the method of sampling can be found in the Chapter 3 (see also Stafford (2017)). Soil types represented by the local set were both Allophanic and non-Allophanic soils with the dominant soil samples (Allophanic 57% and Pumice 16%) not represented in the regional SSL (Fig. 5.1a; Appendix 5.1). The total soil Cd concentration range for the local set was from 0.10 to 2.03 mg Cd/kg with an average of 0.58 mg Cd/kg (Fig. 5.1b; Appendix 5.1).

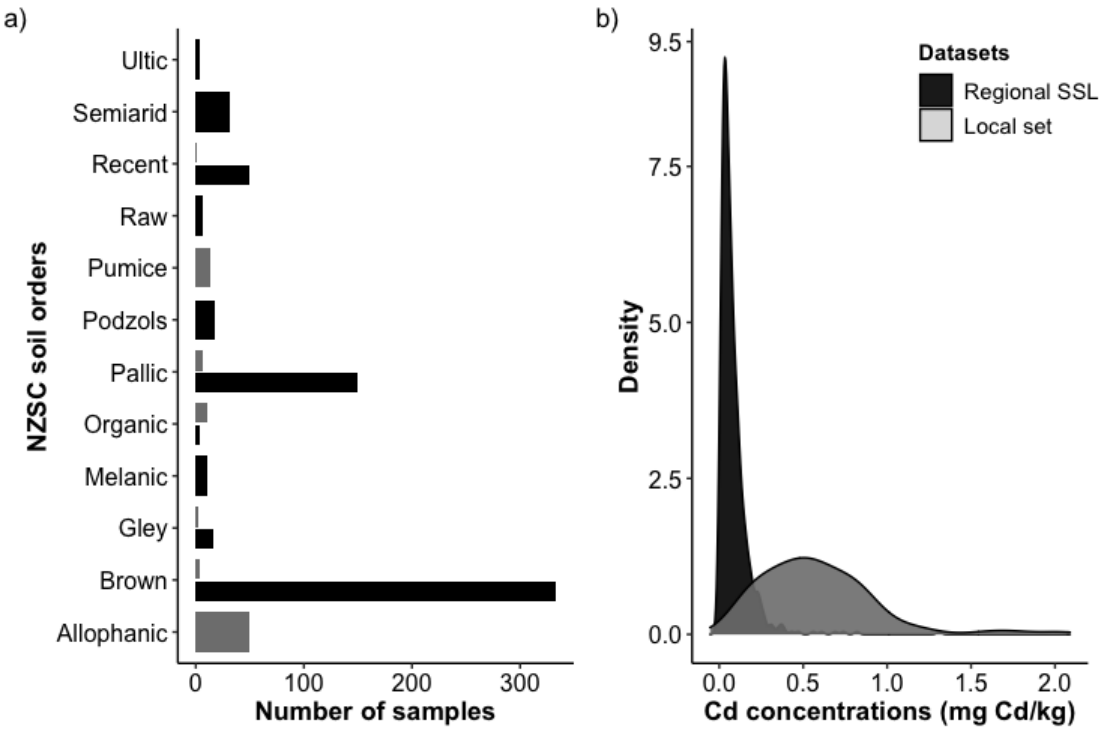


Fig. 5.1 a) Number of samples included in the NZSC soil orders following Hewitt (2010) and b) Cd concentration density plot of the regional SSL (black) and local set (grey).

5.2.1.1 Spectral data collection and pre-processing

The soil sample preparation, spectral data collection, and pre-processing followed the methods detailed in Chapters 3 and 4. For both the regional SSL and local set (Fig. 5.2), vis-NIR (350−2500 nm) reflectance spectra of the prepared samples (air-dried, sub-2 mm

soil) were collected using an ASD FieldSpec 3 spectroradiometer (Analytical Spectral Devices Inc., Boulder, Colorado, USA) fitted with a contact probe containing a 4.5 W halogen bulb as a light source (Fig. 2.4j). Pre-processing of the vis-NIR spectra included splice correction, discarding the noisiest part of the spectra (350−399 nm). The spectral data were converted to pseudo-absorbance ($\log_{10}(R)$) and finally derived and smoothed

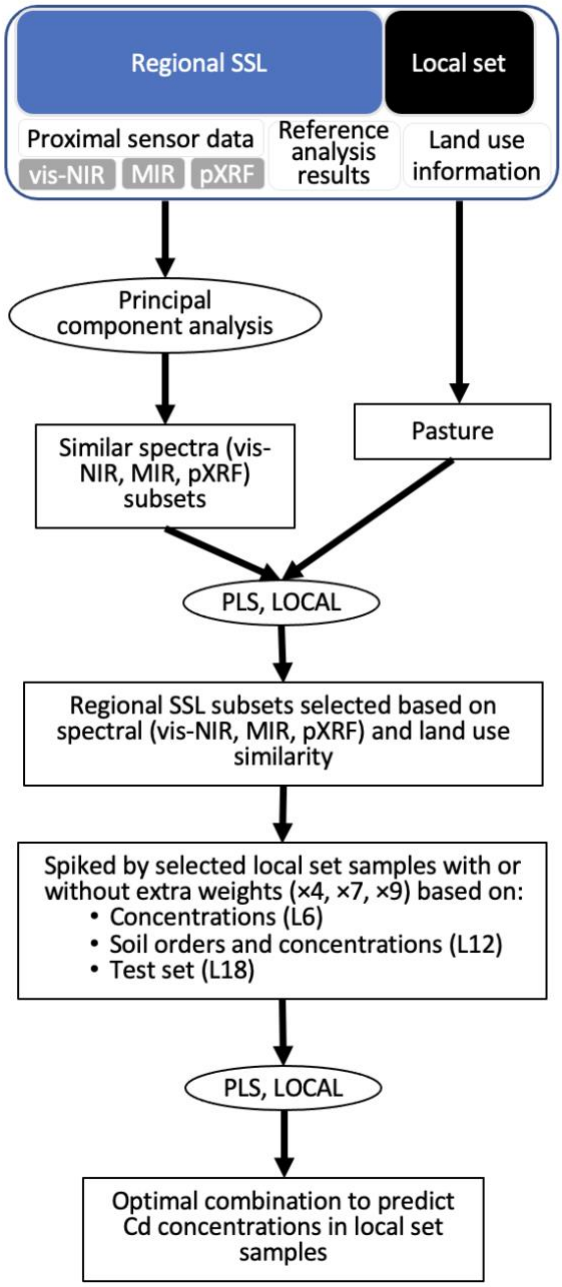


Fig. 5.2 A workflow diagram of the strategy applied to quantify Cd concentration in local samples using the regional SSL in the study.

using a Savitzky-Golay filter (first-order derivative, window size nine, and second-order polynomial) (Savitzky and Golay, 1964).

Mid-infrared (7498−600 cm^{-1}) diffuse reflectance spectra of the prepared samples (air-dried and milled using a Retsch RM200 mortar grinder for 30 seconds) were collected using a 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) (Fig. 2.4n). Raw MIR spectra were derived and smoothed using a Savitzky-Golay filter (first-order derivative, window size nine, and second-order polynomial) (Savitzky and Golay, 1964).

The X-ray fluorescence spectra (0−40 keV) of the prepared samples were collected using an Olympus Vanta C series pXRF instrument containing rhodium anode in *Geochem* mode (*2 beams*) (Fig. 2.4e). Air-dried and sieved to sub 2 mm samples were packed into a cylindrical plastic cup 22 mm height by 14 mm diameter and covered with a 4 µm thick polypropylene transparent film). For this study, *beam 1* raw spectra were used after removing the first 24 wavelength data with zero spectral response values. The pXRF instrument has a built-in detection limit of 5 mg Cd/kg soil (Table 2.6).

The principal component analysis (PCA) of spectral data was performed to reduce the multi-dimensionality of spectral data into a few principal components. Spectral data were centred before performing PCA. The distribution of the regional SSL and local set samples in the principal components space were plotted using the first two principal components describing the greatest amount of variance in the datasets.

5.2.2 Developing a strategy to quantify Cd concentration in local samples using a regional SSL

5.2.2.1 Model development

All chemometric analysis was performed in the R statistical environment (RStudio Team, 2021). Three proximal sensing techniques: vis-NIR, MIR, and pXRF sensor data were independently used as input in the algorithms tested (Fig. 5.2). Total Cd concentration values were log-transformed and then mean centring and variance scaling was performed before use in the predictive modelling. Prediction models were developed using either

PLS or LOCAL algorithms (Fig. 5.2). The R package *pls* (Mevik et al., 2020) was used for PLS regression. Hyper-parametrisation of PLS-based prediction models were carried out using cross-validation with 10-fold splits and 25 repeats using the R package *caret* (Kuhn et al., 2021). The optimal set of parameters chosen was based on the simplest model within the 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 with 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. The PLS was used to select unique samples to use as predictors and validated by the nearest neighbour. Multiple PLS cross-validation (i.e., between a minimum of four and a maximum of 25 components) were used to develop PLS models and predict, which were then weighted and averaged to obtain the final predicted value (Shenk et al., 1997).

5.2.2.2 Selection of samples

For prediction model development, regional SSL subsets were selected based on spectral similarity and land use (Fig. 5.2). The SSL samples spectrally similar to local samples were selected using PCA of each sensor data (vis-NIR, MIR, and pXRF) from the SSL and local set. Spectral data were centred before performing PCA. The distances of a SSL sample from the local set were summed. A SSL sample with the least added distance value was considered the most similar. Consequently, the regional SSL (R625) samples were clustered into R200, R250, R300, R350, R400, R450, R500, and R550 (Fig. 5.3). Based on pastoral use, 283 samples in the SSL were selected and termed RP283 (Fig. 5.3).

5.2.2.3 Spiking

Regional SSL subsets selected based on spectral and land use similarity were independently spiked using selected local samples (Fig. 5.2). A test set (20%) of 18 samples selected from the local set using the Kennard-Stone sampling method were included in the SSL-based prediction model to check improvement in the predictive accuracy (Fig. 5.2). Three selection methods were implemented to assess the test set: (1)

samples containing Cd concentration of 0.60 mg/kg or more to fill the Cd concentration gap in the SSL (L6); (2) Allophanic and Pumice soil orders samples (L12) to fill the soil order gap in the SSL; and (3) the test set (L18). In addition, test set samples used for spiking were given weights of four, seven, and nine times and checked for improvement in the accuracy without weight (Greenberg et al., 2022).

5.2.2.4 Model performance assessment

The accuracy of prediction models was calculated and compared using performance statistics including RMSE, R^2, RPIQ (Bellon-Maurel et al., 2010), CCC (Lin, 1989), and bias. Details on performance statistics can be found in Section 2.4.5 of Chapter 2. These parameters were calculated using the R package *spectacles* (Roudier, 2021).

5.2 Results

5.2.2 Spectral similarity

The PCA of vis-NIR, MIR, and pXRF spectra for the first two principal components show the spread of samples in the regional SSL and local set as per spectra characteristics (Fig. 5.3). The PCA of spectral data showed similarity among samples in the SSL and the local set decreased in the order vis-NIR, followed by MIR, and pXRF (Fig. 5.3). The first two principal components of the pXRF spectra explained 94% spectral variance, with this value reducing to 69% and 68% for the MIR and vis-NIR spectra, respectively (Fig. 5.3). Both the large SSL and local set were in proximity in the principal component space of MIR spectra than vis-NIR and pXRF (Fig. 5.3).

5.2.3 Prediction of Cd locally using a subset from the regional SSL

Regional SSL subsets selected based on the similarity of spectra and land use were employed to predict the Cd concentration in the local set samples, aiming for improved accuracy (Figs 5.4 and 5.6; Appendix 5.1). A Cd prediction model based on the selection of 550 samples from the large SSL (R550 from vis-NIR PCA; Fig. 5.4) as input for

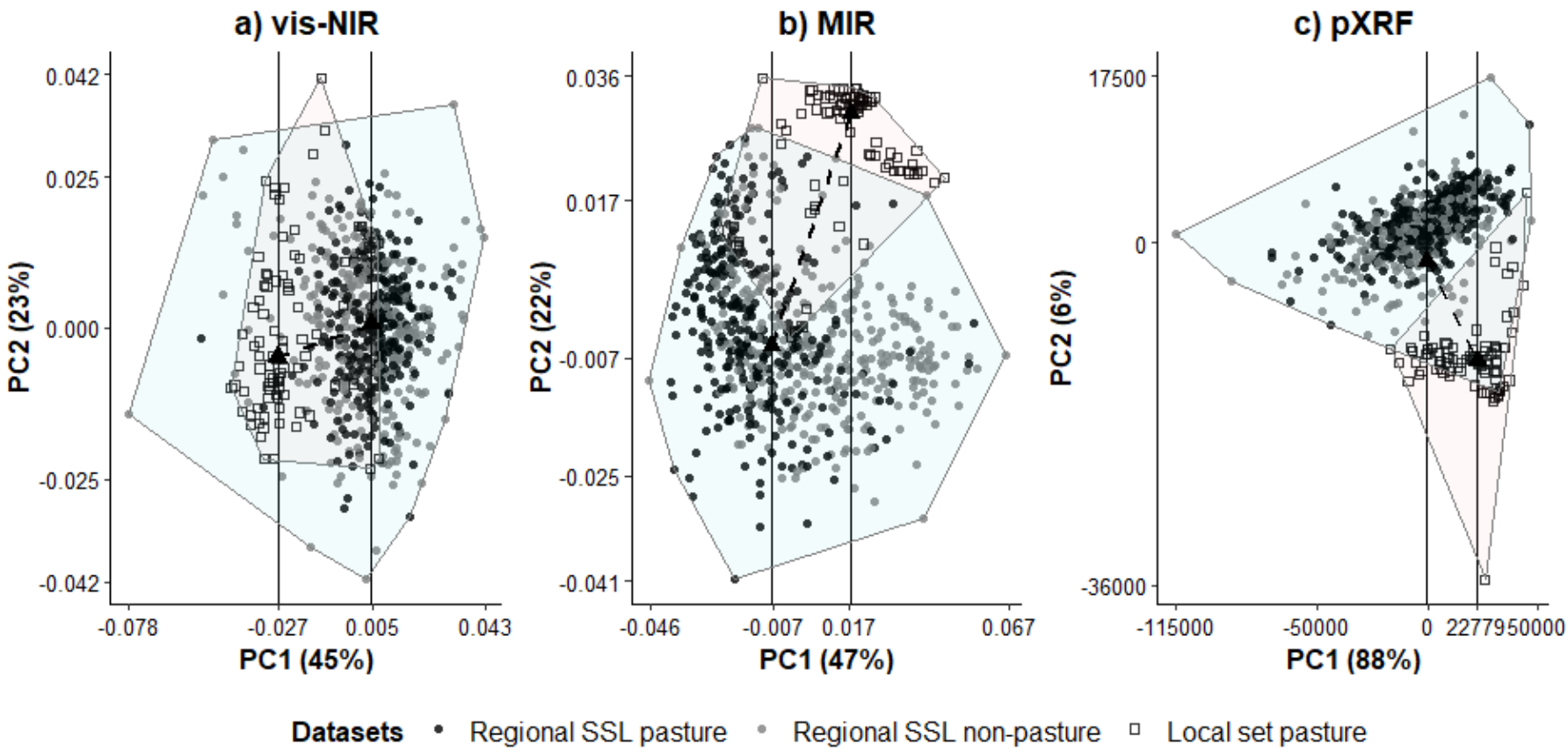


Fig. 5.3 Principal component analysis (PC1 and PC2) of (a) vis-NIR, (b) MIR, and (c) pXRF spectra for the regional SSL (n=625) including pasture (black dots) and non-pasture (grey dots) and the local set pasture (n=87; rectangles) samples. Spectral variance explained by each principal component is shown in % value inside the bracket. Distance between two datasets is indicated by a black dashed line joining centroids (black triangles) of each dataset.

LOCAL algorithm performed with reasonable accuracy, showing a relatively low RMSE (0.37 mg Cd/kg), relatively high CCC (0.53), and RPIQ (1.16) (Fig 5.4a, Appendix 5.2). Independent of the sensors considered, models based on a lower number of samples (e.g., 250, 350) performed with poor accuracy in general while predicting soil Cd locally (Fig. 5.4; Appendix 5.2). Prediction models using sensor data of pastoral soil subset (RP283) as input for PLS or LOCAL algorithms performed with similar accuracy (RMSE of 0.38–0.59 mg Cd/kg) than of spectrally similar subsets (Fig. 5.4; Appendix 5.2). Among the three proximal sensors, model using MIR data of RP283 as input for PLS performed optimal with an RMSE of 0.38 mg Cd/kg, CCC of 0.21, and RPIQ of 1.13 (Fig 5.4; Appendix 5.2).

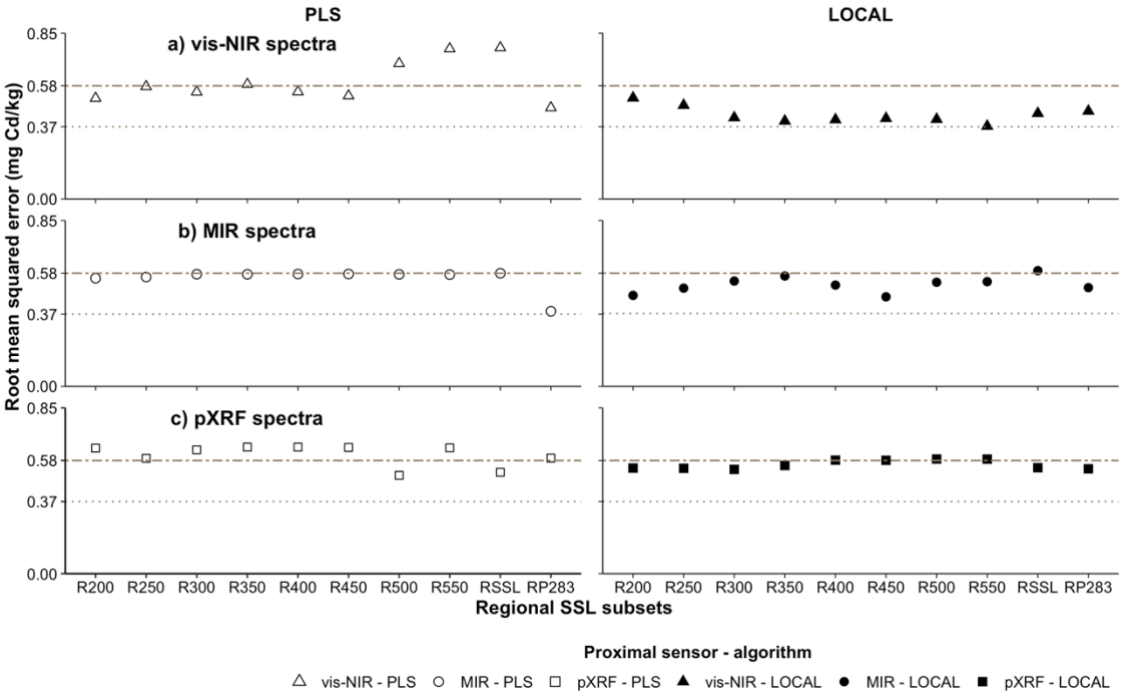


Fig. 5.4 Root mean square error (mg Cd/kg) of Cd concentration for local samples (n=87) predicted using independently three proximal sensors: vis-NIR (triangles), MIR (circles), and pXRF (rectangles) data of regional SSL subsets as input for PLS (hollow shapes) and LOCAL (black coloured shapes) algorithms. Regional SSL subsets (R200, R250, R300, R400, R450, R500, and R550) were selected after PCA analysis of individual sensor data of the regional SSL and local set. The regional SSL (R625) and the SSL pastoral soil subset (RP283) were included for comparison. The dashed line shows the mean Cd concentration (0.58 mg Cd/kg) of the local set.

5.2.4 Predictions of Cd locally by spiking regional SSL subsets

Soil Cd prediction models based on spiked large SSL subsets R450 (from MIR PCA) and RP283 MIR data as input for LOCAL algorithm performed reasonably well with

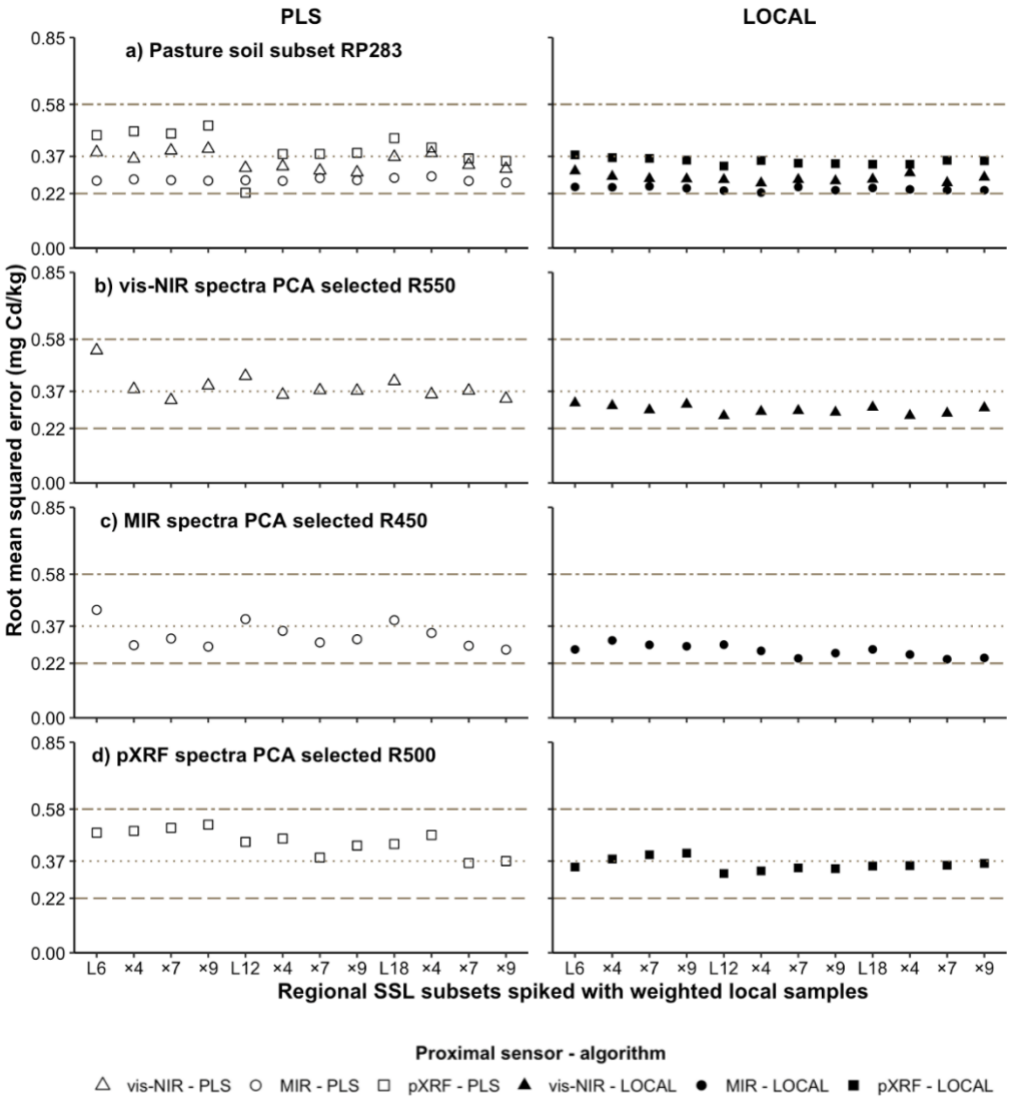


Fig. 5.5 Root mean square error (mg Cd/kg) of Cd concentration predicted for local samples (n=87) using each of the three proximal sensor data: vis-NIR (triangles), MIR (circles), and pXRF (rectangles) using regional SSL subsets a) pastoral soil RP283, b) vis-NIR spectra PCA selected R550, c) MIR spectra PCA selected R450, and d) pXRF spectra PCA selected R500 spiked with extra weighted (×4, ×7, ×9) local SSL set samples (L6, L12, L18) as input for PLS (white coloured shapes) and LOCAL (black coloured shapes) algorithms.

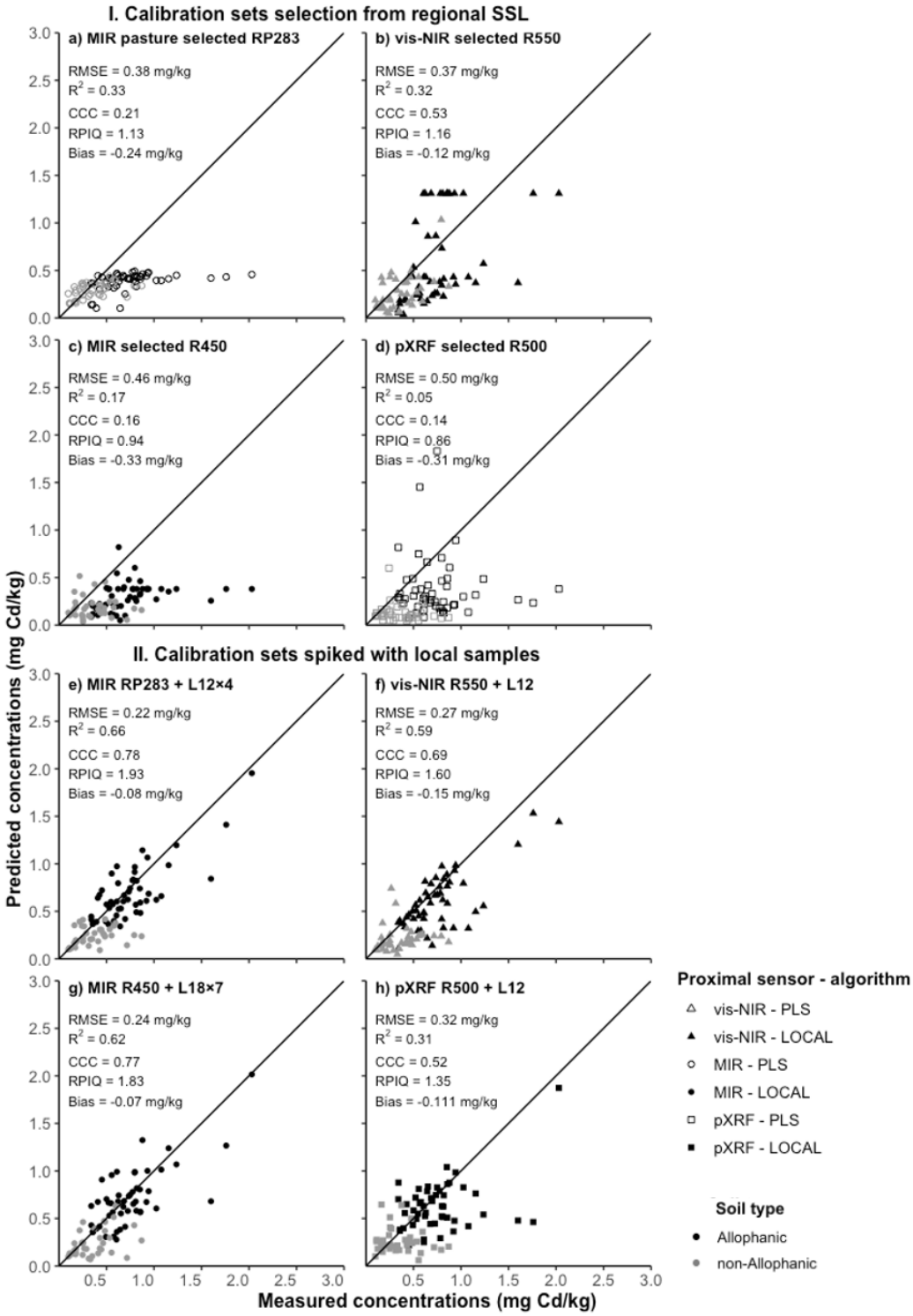


Fig. 5.6 Measured versus predicted Cd concentration (mg Cd/kg) for the local set (n=87) including Allophanic (black) and non-Allophanic (grey) samples based on optimal calibration models using individual proximal sensor data: vis-NIR (triangle), MIR (circle), or pXRF (rectangle) from 1) the regional SSL pastoral soil subset (a) and spectrally similar subsets (b, c, d) and 2) spiked regional SSL subsets as input for PLS (hollow shapes) or LOCAL (black coloured shapes) algorithms.

relatively low RMSE values of 0.22−0.31 mg Cd/kg (Fig. 5.5c; Appendix 5.3). Prediction models based on spiking the large SSL subsets R550 (from vis-NIR PCA) and RP283 vis-NIR data as input for LOCAL algorithm performed a moderate accuracy with RMSE values of 0.26–0.32 mg Cd/kg (Figs 5.5a and 5.5b; Appendix 5.3). Models based on spiking the large SSL subsets R500 (from pXRF PCA) and RP283 pXRF data as input for LOCAL algorithm showed poor performance, with relatively high RMSE values of 0.32–0.40 mg Cd/kg (Figs 5.5a and 5.5d). The LOCAL algorithm outperformed PLS when spiked SSL subsets were used (Fig. 5.5; Appendix 5.3). It is worth noting that models based on LOCAL algorithm performed much more evenly regardless of the weights (RMSE variation of 5%), while models based on PLS regression algorithm were severely influenced by the weights (RMSE variation of 11%) (Fig. 5.5; Appendix 5.3).

Prediction models quantifying Cd using MIR data from spiked SSL subsets outperformed models using vis-NIR and pXRF data (Fig 5.e; Appendix 5.3). Mid-IR data from RP283+L12×4 as input for LOCAL algorithm performed optimal with an RMSE of 0.22, CCC of 0.78, and RPIQ of 1.93 (Figs 5.5 and 5.6e). Spiked SSL subsets including one or more high Cd concentration local samples (0.6–2.0 mg Cd/kg) and filling the soil orders gap improved the accuracy of local sample prediction (Figs 5.6e−h).

5.2.5 Important wavelengths to predict Cd locally using regional SSL subsets, and spiked regional SSL subsets

The PLS loadings highlighted the relative importance of certain spectral regions, and these areas contributed to explain the performance of a particular prediction model developed using the regional SSL, local set, regional SSL subsets, and spiked regional SSL subsets (Fig. 5.7). For Cd prediction using vis-NIR data, loadings showed the importance of the 1300–1450 nm, 1800–2000 nm, and 2200–2500 nm regions (Fig. 5.7a), which corresponds to stretching and bending of molecular OH, H_2O, OH bonds to Al–, Mg– and Fe– containing minerals and soil organic matter (Soriano-Disla et al., 2014). The general pattern of PLS loadings matched between local set and the regional SSL, while importance of the 400–850 nm region (containing organic matter and Fe oxides) differed between these two sets (arrow; Fig. 5.7a).

For Cd prediction using MIR data, loadings showed the importance of the 3700–3000 cm^{-1} and 2130–1700 cm^{-1} (Figs 5.7b and 5.7c; Appendix 2.3), which corresponds to

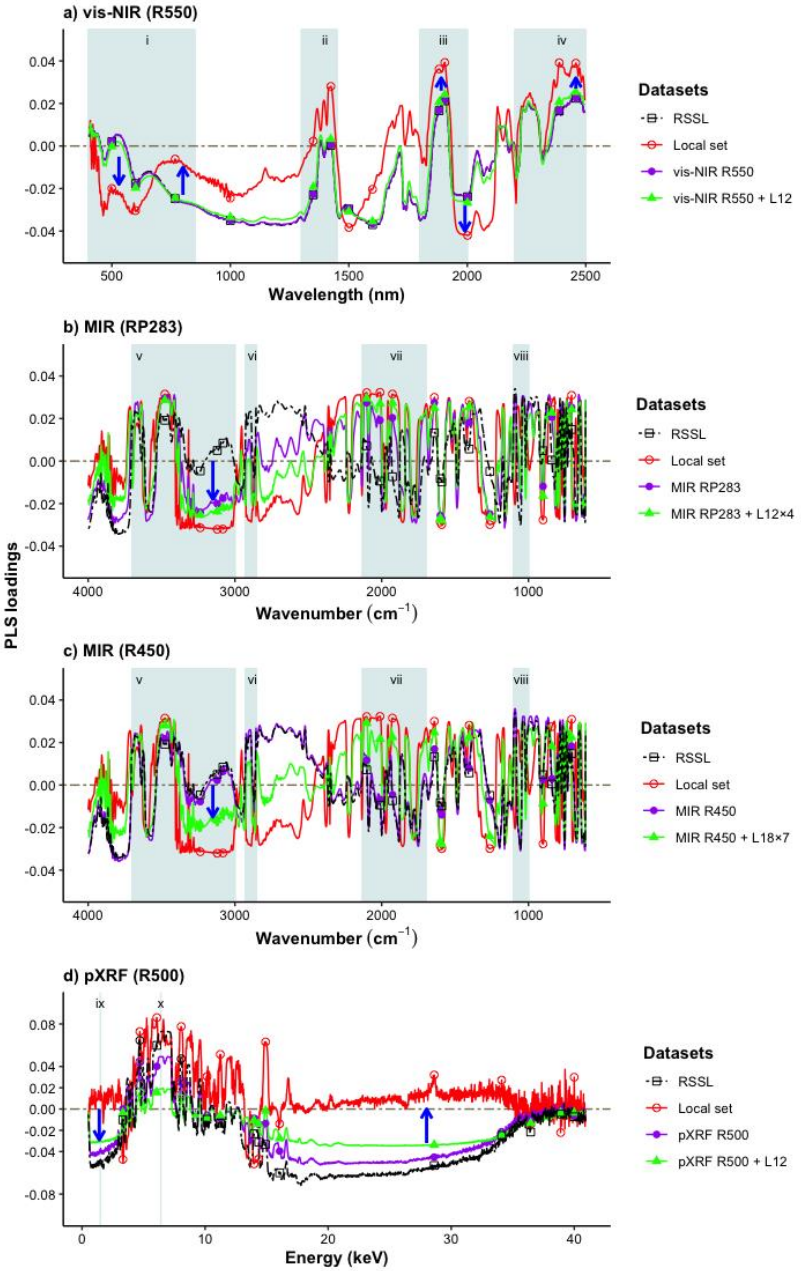


Fig. 5.7 PLS loadings of Cd concentration and spectral response for the regional soil spectral library (RSSL), local set, and selected regional SSL subsets a) R550, b) RP283, c) R450, and d) R500 and their optimal performing spiked sets. The shaded areas in the vis-NIR region: i) soil colour, iron oxides, and soil organic matter (400−850 nm), ii) minerals (1300−1450 nm), iii) Al– and Fe– containing minerals and soil organic matter (1800−2000 nm), and iv) Al– and Fe– containing minerals and soil organic matter (2200−2500 nm); in the MIR region: v) Fe– and Al– containing minerals (3700−3000 cm^{-1}), vi) alkyl (2929−2855 cm^{-1}), vii) metal–carbonyl (2130−1700 cm^{-1}), and viii) quartz (1100−1000 cm^{-1}); and in the XRF region: ix) Al (1.48 keV) and x) Fe (6.40 keV). Arrows show the regions where PLS loadings for the regional SSL and local set differ.

fundamental vibrations of metal–carbonyl bonds, and the OH group of Al– and Fe–containing minerals and H_2O respectively (Wang et al., 2017; Zimmermann et al., 2007). The importance of the 3700–3000 cm^{-1} (Al– and Fe–containing minerals region) and 2130–1700 cm^{-1} (metal–carbonyl region) regions were dissimilar between the regional SSL and local set (arrow; Figs 5.7b and 5.7c).

For Cd prediction using pXRF data, loadings showed the importance of 1.48 keV and 6.40 keV regions, corresponding to Al and Fe respectively (Fig. 5.7d). The PLS loadings for the local set were different to the regional SSL set for both Al and Cd (arrow; Fig. 5.7d). Spiked regional SSL subsets were more similar in PLS loadings to the local set for MIR spectra than vis-NIR and pXRF (Fig. 5.7).

5.3 Discussion

5.3.2 Applicability of the regional SSL for local prediction of Cd concentration

For the first time, this study used a SSL with data from three proximal sensors (vis-NIR, MIR, and pXRF spectroscopy) to rapidly quantify total soil Cd in samples obtained locally (agricultural soils). The accuracy achieved in this study to quantify total soil Cd which has no direct response to reflectance spectra (Kooistra et al., 2001; Wu et al., 2010) was reasonable (Fig. 5.6e). Proximal sensors have generally been restricted to measure soil properties those have spectral response in the reflectance spectra, including soil carbon, nitrogen, and clay content (Nocita et al., 2014; Stenberg and Rossel, 2010; Wu et al., 2010).

The optimal prediction model developed using MIR data as input from the regional SSL pastoral soil subset (RP283) spiked with weighed local samples (Fig. 5.6e) could be implemented for environmental monitoring, such as the TFMS designed to monitor and manage Cd concentration in New Zealand agricultural soils (Gray and Cavanagh, 2022). Such approach could reduce the analytical cost to the farmers and allow intensive spatial and temporal monitoring of pastoral farms based on spectral analysis only (Shepherd and Walsh, 2002). In the current study, the optimal model included 12 out of 87 (14%) local samples analysed by the reference laboratory method, while models using proximal

sensor data were implemented to analyse the remaining samples. Given that a large SSL is available, the increase in analytical cost can be considered moderate.

5.3.3 Role of spiking in model performance

The prediction accuracy of models based on the regional SSL with or without adding local samples (spiking) to quantify Cd in local samples relied mainly on the similarity of (1) spectra distribution (Gogé et al., 2014) in the principal component space (Fig. 5.3), and (2) association of Cd with spectrally active soil components including soil organic matter and Al– and Fe–containing minerals (Ng et al., 2022a) as observed in the PLS loadings (Fig. 5.7).

When the prediction accuracy of models based on spiked SSL subsets were compared, those models developed using MIR data were relatively accurate in estimating Cd locally (Figs 5.5 and 5.6; Appendix 5.3; Sanderman et al. (2021)). This quantification was possible due to the observed similarity in the PLS loadings of MIR spectra between the spiked SSL pasture subset and the local set (Fig. 5.7b) indicating that Cd in both sets were associated with Al– and Fe–containing minerals (3700–3000 cm^{-1}) and metal–carbonyl groups (2100–1730 cm^{-1}) (Appendix 2.3; Janik et al. (1998); Niazi et al. (2015); Wang et al. (2017)).

Noise from overtones captured by the vis-NIR spectra from the fundamental vibrations of molecular bonds and functional groups (Siebielec et al., 2004; Stenberg and Rossel, 2010) limited the accuracy of prediction models using this sensor data (Figs 5.6e and 5.6f; Appendix 5.3). In general, the poor performance of models using pXRF data to quantify Cd in local samples may be mainly related to the low Cd concentration in the regional SSL samples (average=0.08 mg. Cd/kg soil; Fig. 5.1b) (Lemière, 2018; Weindorf and Chakraborty, 2020).

The moderated success in predicting Cd achieved by the selection method may have been influenced by the limited comprehensiveness of the regional SSL used in this study (Figs 5.1, 5.4, and 5.6a-d; Appendix 5.1; Gomez et al. (2020)). The regional SSL in the current study is dominated by Brown soils (53%) order, the absence of Allophanic soils, while the Cd concentration distribution is skewed to the right with mostly less than 0.5 mg Cd/kg soil, and contrasted land-use types, with more than 50% samples represent non-

pasture land use (Chapter 3; Appendix 5.1). The "bias" found in the regional SSL may have caused poor predictive accuracy due to: (1) the difference in spectral similarity of soil types to Allophanic and Pumice soil present in the local set (Fig. 4.5; Chapter 4; Asgari et al. (2020)), (2) difference in the spectral characteristics of non-pastoral samples in the regional SSL to only pastoral samples in the local set (Fig. 5.3), and (3) soil depth difference between the local set (0−15 cm) and the SSL (0−20 cm) which might have created noise to predict Cd using the regional SSL (Moura-Bueno et al., 2020).

Selecting representative samples with common land use in both the regional SSL and local set increased the success when predicting Cd locally up to a point (i.e., RP283; Fig. 5.6a). This finding is in line with Moura-Bueno et al. (2020) who found that land use-based selection of a vis-NIR SSL subset for soil organic carbon prediction in local samples outperformed other selection methods such as geographic regions, soil texture class, and spectral similarity. This selection method based on land use (up to a moderate number of samples of 283) promoted that the samples included later in the model were (1) spectrally similar and (2) the range of Cd concentration was comparable (Figs 5.1, 5.3, and 5.7b) as shown by Lobsey et al. (2017) for local carbon prediction using the spectral library.

Selection based on spectral similarity alone (e.g., using PCA) may allow a moderately accurate prediction of Cd (Fig. 5.4). Nawar and Mouazen (2018) also described superior accuracy when predicting soil carbon locally (based on the vis-NIR sensor) by analysing similarity with PCA. Principal component analysis selected a subset of samples contributing to predict soil carbon, while the Kennard-stone sampling and random selection methods were less successful in their study (Nawar and Mouazen, 2018).

Spiking (particularly in combination with selection and weighing: RP283+L12×4; Figs 5.6 and 5.7) provided an additional advantage when predicting Cd locally, improving the accuracy of prediction models based on MIR data from SSL subsets with local samples by 27% (Fig. 5.6; Appendix 5.1). The improvement in Cd prediction may be caused by spiking the SSL subset with 12 local samples including Allophanic and Pumice soils and ≥0.6 mg Cd/kg, thus filling key gaps in the database (pedodiversity and concentration) (Brown, 2007; Guerrero et al., 2010; Wetterlind and Stenberg, 2010).

Regarding the algorithms considered in this study, LOCAL and PLS, LOCAL was robust while PLS was sensitive to spiking and weights (Figs 5.5 and 5.6; Appendices 5.1 and

5.2). Thus, LOCAL algorithm should be preferred for consistency. The LOCAL algorithm selects similar but unique sample spectra and discards duplicates (Shenk et al., 1997). LOCAL algorithm approach of selecting similar spectra from the SSL and developing a local model specific to the dataset showed better performance than PLS (Dangal et al., 2019). In contrast, PLS selects latent variables based on co-variance in the dataset showing improved accuracy with higher weights for spiked samples (Greenberg et al., 2022; Jiang et al., 2017). Quantitative representation with spiking and weights in PLS improved Cd prediction but did not reach the same accuracy as spiking without weights in LOCAL algorithm (Figs 5.5 and 5.6; Appendices 5.1 and 5.2).

5.3.4 Practical implications and limitations of SSLs

National level soil monitoring can benefit from building a national SSL, which could contribute to a cost-effective assessment of soil properties including Cd concentration at the local scale using proximal sensors (Ng et al., 2022a). The SSLs developed by institutions and countries available in the public domains (e.g., Mendes et al. (2022); Viscarra Rossel et al. (2016)) using sensors with different specifications may require standardisation before they can be applied to different setups (also known as calibration transfer) to quantify soil properties (Pittaki-Chrysodonta et al., 2021; Sanderman et al., 2021). The development of (pedo-)transfer functions correcting field variations of soil moisture, texture, and matrix effect for each soil type are essential to use the SSLs to monitor local soil properties on the go including airborne and spaceborne (hyperspectral) remote sensing (Kuang and Mouazen, 2013; Minasny et al., 2011; Wang et al., 2022).

5.4 Conclusions

Soil spectral libraries could potentially be used to quantify total soil Cd in local samples to support Cd monitoring. This study developed a strategy to customise a prediction model using individual proximal sensor data that could accurately quantify Cd at a farm scale while leveraging a soil spectral library containing data from three proximal sensors. Overall, models using MIR sensor data as input for LOCAL algorithm outperformed vis-NIR and pXRF to quantify Cd at low concentration (≤2 mg Cd/kg). This study provided a mechanistic explanation of accurate Cd quantification through SSL subset selection and spiking using PLS loadings. Mid-IR sensor quantified Cd associated with similar

spectrally active soil components in the regional SSL and local set for accurate prediction. Selection of SSL subset with similar land use to local samples improved total soil Cd prediction. In the scenario where there is a lack of land use information, a SSL subset with a similarity of spectra in the principal component space to local samples can be useful for predictive modelling. For SSLs which are limited in coverage of soil orders and Cd concentration, spiking the SSL subset using local samples that "fill the gap" can improve prediction accuracy. Using spectral data, the Kennard-Stone sampling method can efficiently select local samples to be included in the SSL-based model. Developing SSL at the national scale covering all pedogenic variations, land uses, and soil properties of concern can make local sample characterisation inexpensive and rapid for data driven decision making processes.

Chapter 6

Analysis of reflectance and fluorescence spectra to assess cadmium in soil and plant: glasshouse experiments with chicory (*Cichorium intybus* L.)

Abstract

Regular soil Cd assessment has been implemented as a part of the Cd monitoring strategy in New Zealand to reduce the environmental, economic, and health risks of Cd accumulation in agricultural soils. Proximal sensing techniques including vis-NIR, MIR, and pXRF spectroscopy are increasingly considered as an inexpensive complement to reference laboratory methods. Using proximal sensors to estimate Cd in plants, along with Cd in soils, will help to monitor Cd accumulation and translocation by plants. Two independent glasshouse experiments were conducted using Pallic soil (experiment I) and Allophanic soil (experiment II) amended with increasing Cd concentrations (0, 0.5, 0.75, 1.0, 2.0, 3.5, and 5 mg Cd/kg soil), with and without chicory (*Cichorium intybus* L.) plants with four replicates of each treatment combination. Soil and plant aboveground biomass were sampled periodically and plant roots were harvested at the end of each experiment. *In-situ* soil and leaf clip vis-NIR spectra were collected and representative prepared soil and plant samples were scanned using vis-NIR, MIR, and pXRF sensors in the laboratory. Cross-validated prediction models to quantify Cd concentration in soil and plant samples were developed implementing partial least squares regression analysis of pre-processed spectra. Model predictions were used to estimate Cd bioconcentration and translocation factor values for chicory. There was a significant relationship (R^2=0.74, $p<0.001$) between measured and predicted translocation factor values. There was a significant relationship (R^2=0.36; $p<0.001$) between plant above-ground biomass Cd and a blue green index 2 (R_{450}/R_{550}). A model using *in-situ* leaf clip vis-NIR spectra quantified plant above-ground biomass Cd (n=82, 2.40–56.77 mg Cd/kg) reasonably well with $nRMSE_{cv}$ of 28% and CCC_{cv} of 0.93. A model using laboratory vis-NIR spectra predicted well Cd

in Pallic soil (n=224; 0.17–5.45 mg Cd/kg) with $nRMSE_{cv}$ of 22% and CCC_{cv} of 0.97. Models using pXRF spectra assessed Cd concentration in Allophanic soils (n=112, 0.41–4.81 mg Cd/kg) and plant root (n=28, 0.86–25.79 mg Cd/kg) very well with $nRMSE_{cv}$ of 16% and 9% and CCC_{cv} of 0.95 and 0.99, respectively. The quantitative prediction models developed in this study showed potential to rapidly assess soil and plant Cd concentrations and to calculate translocation factor values for evaluating Cd food chain transfer risks.

6.1 Introduction

Plant Cd accumulation from agricultural soils with long-term phosphate fertiliser application history has raised public concern due to the potential health impact of Cd in food product (Morgan, 2010). Soil-to-plant transfer is the major pathway for human and animal exposure to Cd (Järup and Åkesson, 2009). Studies have shown that agricultural soil Cd concentration have increased from background level (Abraham, 2018; Cavanagh et al., 2015; Reiser et al., 2014) and at elevated soil concentrations, phyto-available Cd may also increase (Ubeynarayana et al., 2021). Soil Cd availability for plant uptake is influenced by multiple edaphic and management factors including soil pH, organic matter content, clay minerals, and liming (Christensen and Haung, 1999; Cottenie and Verloo, 1984; Loganathan et al., 2012; Marković et al., 2019).

Plants including forage herb species such as chicory (*Cichorium intybus* L.) accumulate phyto-available Cd from the soil to root and then translocate this TE to aboveground biomass (Abe et al., 2008; McKone and Maddalena, 2007). The Cd bioconcentration factor (BCF) can be defined as a plant's ability to accumulate Cd and is a function of the soil Cd concentration (Zayed et al., 1998), while the Cd translocation factor (TF) can be defined as a plant's ability to translocate root accumulated Cd to aboveground biomass (Mattina et al., 2003). Literature studies have shown that for plants grown in soil containing 0.2 to 11 mg Cd/kg soil, Cd BCF and TF values were found in a wide range (Cd BCF root biomass: 0.9–1011, Cd BCF aboveground biomass: 0.1–453, and Cd TF: 0.3–29) (Abe et al., 2008; Guérin et al., 2022; Stafford et al., 2016; Yang et al., 2021). Consumption of plants with a high Cd BCF or high TF may expose pasture grazing animals to dietary accumulation of Cd (Anderson et al., 2022; Lee et al., 1996).

Proximal sensing techniques including vis-NIR (350–2500 nm), MIR (4000–400 cm^{-1}), and pXRF (0–40 keV) spectroscopy are widely deployed to assess total soil Cd in mining areas and across contaminated agricultural soils (Nawar et al., 2019). The use of proximal sensors to estimate Cd concentration in plant samples is limited to concentration inducing stress or toxicity in the plant (Artz et al., 2008; Shi et al., 2016; Sridhar et al., 2007). Cadmium assessment in soil using reflectance spectroscopy is linked to the association of Cd with spectrally active soil components including soil organic matter and Al–, Fe–, and Mn–containing minerals (Soriano-Disla et al., 2014) whereas spectral information from plants may indicate changes in pigment content, cell structure, and biochemical composition maybe due to Cd accumulation (Rathod et al., 2018; Shi et al., 2016).

Portable-XRF and vis-NIR spectroscopy has been used to estimate Cd at concentration that will induce stress or toxic effect in plants such as *Festuca* sp., *Miscanthus sacchariflorus*, and *Phalaris aurundinacea* (Feng et al., 2019; Götze et al., 2010; McGladdery et al., 2018). Mid-IR spectroscopy has been applied to assess plant composition, plasticity, and disease as this technique is capable of distinguishing the principal chemical classes such as carbohydrates, proteins, and fats through the fundamental vibrational characteristics of their structural chemical bonds (Artz et al., 2008; Largo-Gosens et al., 2014; Zhang et al., 2017). Using pXRF spectra as input for simple linear regression model developed to predict Cd, McGladdery et al. (2018) estimated plant Cd concentration (10–265 mg Cd/kg dry matter) with RMSE of 0.07 mg Cd/kg, R^2 of 0.75, and RPIQ of 3.27. Most of these studies scanned plant materials in laboratory conditions, i.e., dried, ground/sieved samples.

In-situ plant vis-NIR spectra have shown promising results in assessing plant Cd accumulation to toxic concentration (Rathod et al., 2015a; Sridhar et al., 2007). *In-situ* measurement is taken in the field in proximity or contact with the plant. The resulting plant spectra contain information on differences in pigment content, leaf thickness, and biochemical composition that may be due to Cd accumulation (Bandaru et al., 2010; Rosso et al., 2005). Feng et al. (2019) used *in-situ* vis-NIR spectra as input for PLS and developed prediction models to estimate Cd in leaves (up to 800 mg Cd/kg DM) and roots (up to 12000 mg Cd/kg) with R^2 0.87 and 0.88, and RMSE of 92 mg Cd/kg DM and 1640 mg Cd/kg DM, respectively.

Vegetation indices have also been developed using *in-situ* vis-NIR data as indicators of vegetation and soil conditions including toxic Cd concentration (Kooistra et al., 2004; Wu et al., 2019). Vegetation indices are unitless ratios or linear combinations of two or more spectral bands, which enhance plant spectral signal and reduce background effect (Rathod et al., 2013). Vegetation indices have shown a good relationship with Cd at plant toxic concentration (Feng et al., 2019). Further, vegetation index calculated with remote sensing data can be validated using a vegetation index measured using *in-situ* reflectance spectra (Xue and Su, 2017).

In general, TE experiments are conducted to determine bio-toxicity levels with respect to plant growth and development (Jones et al., 2008). Such experiments can also generate knowhow on the concentration effect of target analyte such as Cd in plants (Sridhar et al., 2007). With the advancement of spectroscopy, such experiments have been used to investigate the effect of different concentrations in the spectral characteristics and to develop vis-NIR spectroscopy-based statistical models (Rathod et al., 2015b; Sridhar et al., 2007). Using a concentration gradient above the values of Cd usually found in soils (e.g., Chapters 3 and 4), mathematical modelling becomes easier (Chen et al., 2019), giving the opportunity to develop robust Cd prediction models not easily generated using natural soils. Including plant in such experiments can be used to quantify the difference in plant uptake and spectral characteristics as a function of concentration. The objectives of the research described in this Chapter were to 1) assess the effect of increasing soil Cd concentration on plant biomass and Cd uptake, 2) apply vegetation indices to detect plant Cd concentration rapidly, and 3) to predict the concentration of Cd in soil, plant aboveground biomass, and root using laboratory MIR, pXRF, vis-NIR and *in-situ* vis-NIR spectra. To achieve these objectives, soil and plant samples were taken periodically from two glasshouse experiments set up in contrasting soils (one Pallic and another Allophanic soil) with controlled amendment of increasing Cd concentrations into pots growing chicory plants, were scanned, both in *in-situ* and in the laboratory.

6.2 Materials and methods

6.2.1 Experiment set up and management

Two independent glasshouse experiments (I and II) were conducted during 2020–2021. Topsoils (0–20 cm) were collected from two typical dairy farms with contrasting and characteristic agricultural soil types in New Zealand, one Pallic soil and another Allophanic soil (Hewitt et al., 2021). For experiment I, an Argillic-fragic Perch-gley Tokomaru Pallic Soil (Planosols as per World Reference Base; Aeric Kandiaqualf as per United States Soil taxonomy) was collected from Dairy No. 4 farm of Massey University, Manawatu campus in January 2020. For experiment II, a Typic Orthic Allophanic soil (Andosols as per World Reference Base; Typic Hapludand as per United States Soil taxonomy) was collected from a commercial dairy farm in Stratford, Taranaki in October 2020. The physical and chemical properties of these two soils are summarised in Table 6.1. Cadmium concentration in both soils were within the background levels obtained in New Zealand (Cavanagh et al., 2015). Total soil Cd concentration in the Allophanic soil was higher than in the Pallic soil, and this may be related to (1) the influence of parent material and/or (2) a history of greater amount of phosphate fertiliser input to the Allophanic soil as observed in the total phosphorus concentration for the two soil types (Table 6.1; Chapter 4; Hewitt et al. (2021); Longhurst et al. (2004)).

Table 6.1 Soil physical and chemical properties of Pallic and Allophanic soils (0–20 cm) used in the study.

Soil properties	**Experiment I**	**Experiment II**
NZ soil classification order[1]	Pallic	Allophanic
Physical properties[2]		
Bulk density (g/cm^3)	1.01	0.70
Soil texture	Silt loam	Loam
Chemical properties[2]		
Total Cd (mg/kg)	0.21	0.46
Soil pH	5.30	5.62
Total C (%)	7.62	9.84
Total N (%)	0.26	0.82
Total P (mg/kg)	0.44	1.94
CEC (cmol/kg)	22	17

[1]Hewitt (2010), [2]Blakemore et al. (1987)

At both sites, topsoil was collected, air-dried, and sieved to sub-4 mm. Based on bulk density, 1500 g Pallic soil (experiment I) and 1127 g Allophanic soil (experiment II) was used to fill 2 L plastic pots to 11.8 cm depth with a soil volume of 1485 cm^3 (Table 6.1; Figs 6.1a and 6.1b). Soil was amended with seven increasing Cd concentrations treatments: 0 (control), 0.50, 0.75, 1.0, 2.0, 3.5, and 5.0 mg Cd/kg soil. Cadmium was sprayed evenly to weighed soils using the corresponding volume taken from 0.08 mg Cd/ml solution prepared using cadmium chloride ($CdCl_2.2.5H_2O$) that would give the target soil Cd concentrations (Figs 6.1c, 6.1d, and 6.1e). Following careful soil mixing, the potted soil was watered to 70% field capacity using deionised water. Then pots were kept at 25°C for a 15-day period (Kim, 2006). This process was repeated for two further cycles allowing Cd to adsorb within the soil matrix (Ubeynarayana et al., 2021).

For each experiment, eight replicate pots per Cd treatment were used for a total number of 56 pots. Out of the eight pots per Cd treatment, four were planted with chicory (*Cichorium intybus* L. cultivar "Puna II") and four were left as unplanted control (Fig. 6.1l). Additionally, one unplanted control pot for each Cd concentration treatment was maintained to backfill the soil sampling spot. Chicory seedlings were initially grown in a controlled temperature room at 21°C with a 12-hour photoperiod. Two seeds were placed near the surface of each hole of the seedling tray filled with propagation sand. Water content was maintained manually to keep the propagation sand moist. Two chicory seedlings each in the appropriate pots were transplanted after 17 days of seed sowing on 22 June 2020 and 19 March 2021 for experiments I and II, respectively (Fig. 6.2). All pots were watered regularly to reach 70% water-holding capacity. Pots were kept inside the controlled temperature room at 21°C with a 24-hour photoperiod for a week after transplanting to ensure establishment. After a week, one seedling was removed from each pot and pots were transferred to a glasshouse.

Both experiments were carried out in the glasshouse conditions at a mean day/night temperature of 19°C (15–28°C) over the length of both experiments. Lights were used to maintain a 12-hour photoperiod. All pots were laid out in a completely randomised design and rotated weekly. Soil water content was maintained at up to 70% water-holding capacity throughout the two experiments by regular watering. No mineral fertilisers were added. Experiments were carried out 18 weeks after transplanting (experiment I, 127 days and experiment II, 132 days) (Fig. 6.2).

6.2.2 Soil sampling, aboveground biomass harvest, and root separation

In both experiments, soil samples were taken from each pot, using a corer (2.80 cm diameter, 11.8 cm depth) to extract a soil column (Fig. 6.1g) from one spot. In experiment I, samples were taken on days 0 (following incubation and before seedling transplant), 21, 45, 60, 73, 101, and 127 (Fig. 6.2a). In experiment II, samples were taken on days 0, 21, 45, 60, 97, and 132 (Fig. 6.2b). Each sampling spot was backfilled with the corresponding soil taken from an additional pot (equivalent Cd concentration treatment) maintained at the same experimental conditions. Soil sampling sites were rotated clockwise around each pot, with the backfilled soil sampling location marked on the parallel top edge of the pot with a coloured line. All soil samples were air-dried and sieved sub-2 mm for further analysis.

In both experiments, chicory aboveground biomass was harvested three times after reaching a plant height of 25 cm (Fig. 6.2; Powell et al. (2007)). Aboveground biomass was harvested 7 cm above the soil surface (Fig. 6.1l; Li and Kemp (2005)). Plants regrew after the first and second harvests. The final harvest of aboveground and root biomass was accomplished after the last soil sampling (Fig. 6.1m). In experiment I, aboveground biomass was harvested on days 73, 101, and 127 and root biomass on day 127 (Fig. 6.2a). In experiment II, aboveground biomass was harvested on days 60, 97, and 132, and root biomass on day 132 (Fig. 6.2b).

Roots were cleaned with tap water to remove soil, while aboveground biomass was rinsed with tap water. Aboveground and root biomass were separated (Fig. 6.1m), and oven-dried (60°C for three days) for dry matter yield (g DM per pot) and aboveground biomass and root Cd (mg Cd per kg DM) assessment. Dry matter yield per pot was converted to yield per ha (Equation 1). Dried plant samples were finely ground for further analysis.

Dry matter yield (kg/ha) = Dry matter yield (g/kg soil) x Soil weight (kg/ha). (1)

where, Dry matter yield (g/kg soil) = Dry matter yield (g/pot)/Soil weight (kg/pot)

Soil weight (kg/ha) = bulk density (kg/m^3) x pot soil depth (m) x 10,000 m^2

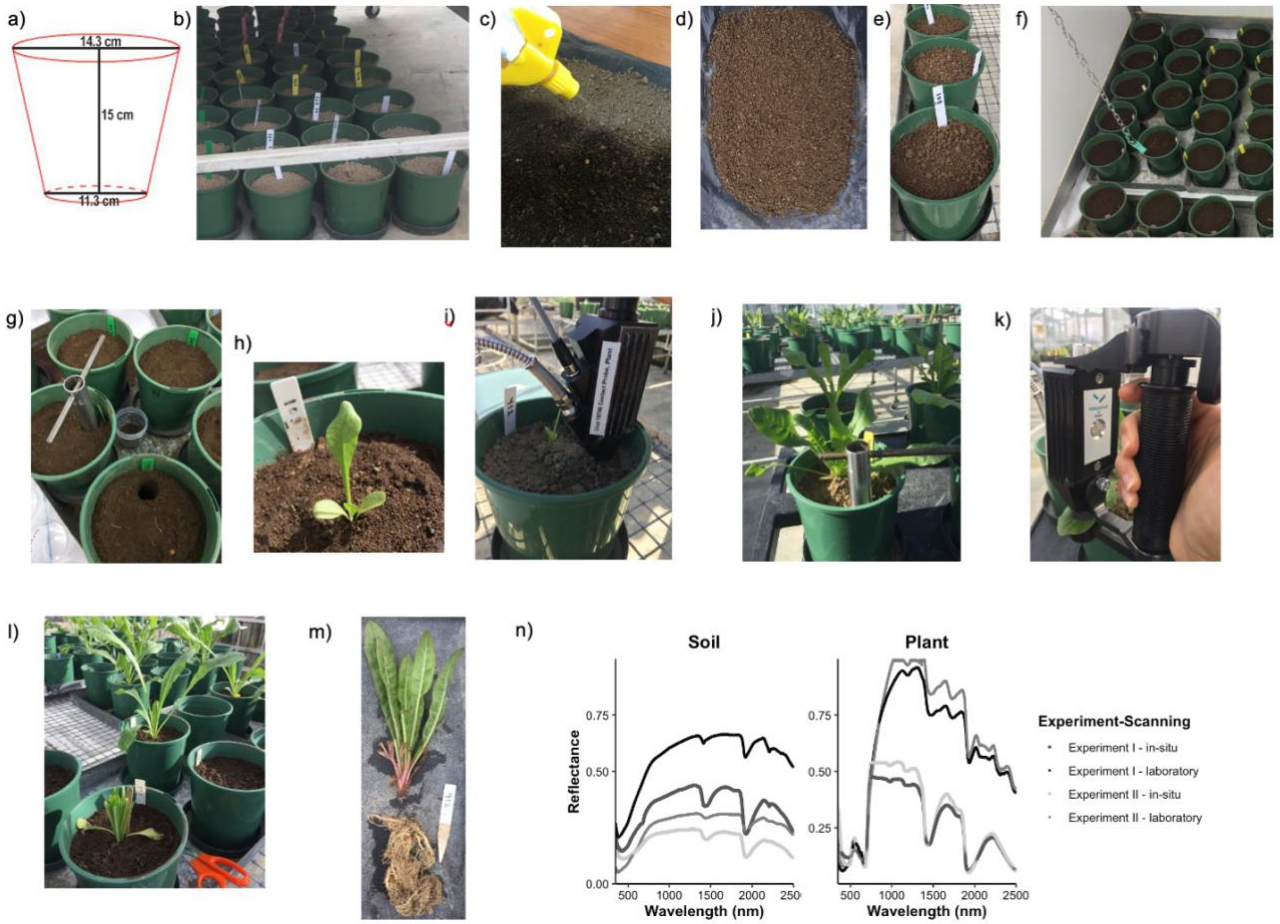
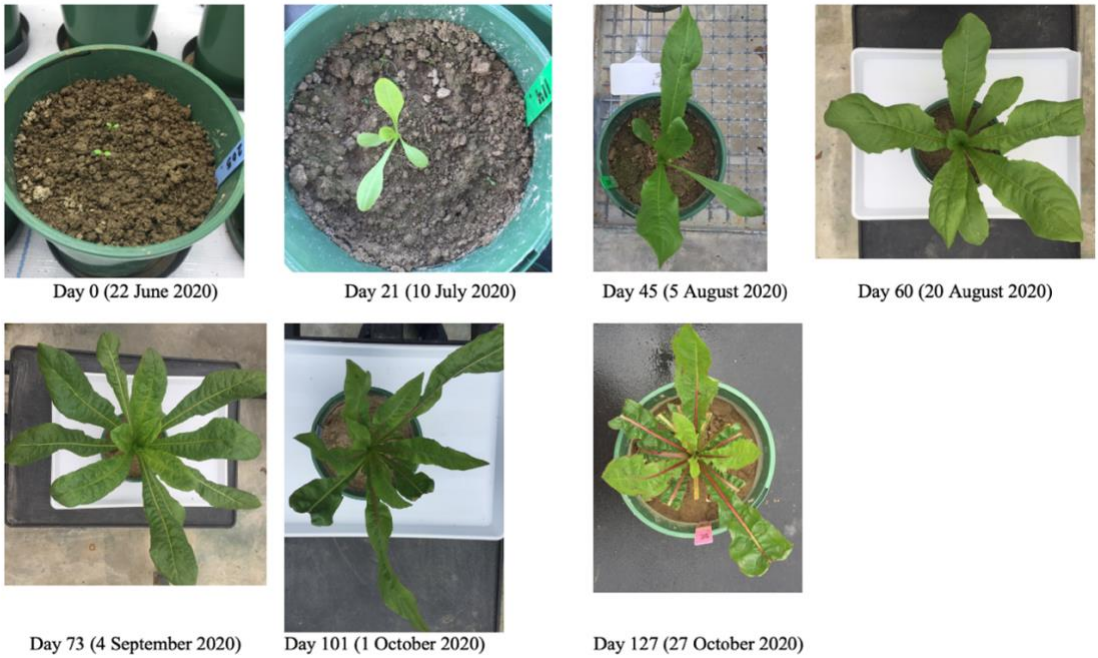


Fig. 6.1 Glasshouse experiments activities and details: a) pot dimensions, b) potted weighed soil, c) spraying Cd solution in the pot soil spread on a plastic sheet, d) soil homogeneously mixing with Cd solution, e) refilling pots with soil amended with Cd solution, f) pot soil incubation, g) a soil core sampling after incubation, h) transplanting chicory seedling in a pot, i) periodic *in-situ* soil scanning, j) periodic non-repeated spot soil core sampling, k) leaf clip scanning, l) aboveground biomass harvest, m) final aboveground biomass and root harvest, and n) *in-situ* and laboratory soil and leaf clip vis-NIR reflectance spectra from experiments I and II.

a) Experiment I



b) Experiment II

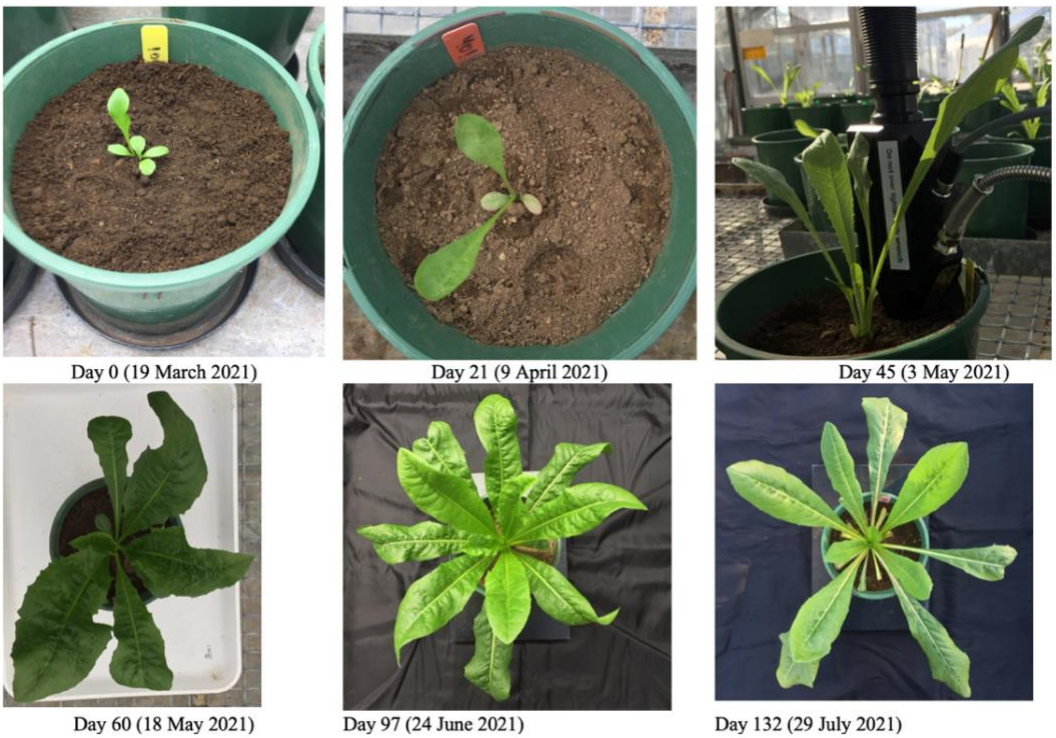


Fig. 6.2 Plant growth photographs with dates aligned with *in-situ* soil scanning, soil core sampling, leaf clip scanning, and aboveground biomass harvest activities in experiments a) I and b) II. Pot diameter was 14.3 cm.

6.2.3 Soil and plant Cd analyses

Soil and plant samples collected from each sampling event were analysed for total Cd concentration. Briefly, total Cd was determined from a sub-2 mm 1 g (soil) or 0.1 g (plant) subsample digested with 10 ml concentrated nitric acid (HNO_3) in the heating block for 2 h (Kovács et al., 2000). The remaining solution was filled to a standard 25 ml volume with 2% HNO_3 before analysing in the graphite furnace atomic absorption spectrometer (GFAAS PinAAcle 900Z, PerkinElmer, UK).

Laboratory Cd analysis was continuously disrupted due to Covid-19 lockdowns and associated workplace and travel restrictions during 2021. Consequently, only a subset of samples (448 soil samples, 110 aboveground biomass samples, and 28 root samples; Appendix 6.1), covering Cd concentration variability, were analysed by GFAAS.

6.2.4 Spectral data collection and pre-processing

6.2.4.1 *In-situ* vis-NIR spectra of soil and plant

At each soil sampling (Fig. 6.2), *in-situ* soil was scanned in three spots in each pot using an ASD FieldSpec 4 spectroradiometer collecting ten vis-NIR spectra from each spot (Figs 6.1i and 6.2n). The instrument was calibrated using a Spectralon white reference panel at the start and after every ten spectra measurements. At each harvest, three newly developed leaves in each pot were scanned by a leaf clip connected to the ASD FieldSpec 4 spectroradiometer through optic fibre, collecting 30 total spectra per pot (Fig. 6.1k).

6.2.4.2 Laboratory scanning of soil, plant aboveground biomass, and root samples

Air-dried soil and fine ground plant subsamples were scanned using the three proximal sensing techniques: vis-NIR, MIR, and pXRF sensors in the laboratory. Details on sensors and scanning can be found in Chapters 2 (Fig. 2.4), 3, and 4. Briefly, an ASD FieldSpec 3 spectroradiometer was used to record a vis-NIR (350–2500 nm) reflectance spectra which was an average of 50 internal readings within the scanning period (Fig. 2.4j). Each sample was scanned in triplicate, rotating petri dish containing sample clockwise 90° angle. A Fourier-transform infrared spectrometer (Vertex 70, Bruker, Germany) equipped

with a microplate reader extension for high throughput screening infrared spectroscopy equipment (HTS-XT, Tensor II, Bruker, Germany) was used to record diffuse MIR (7498–600 cm^{-1}) reflectance spectra of samples (four replicate per sample) (Fig. 2.4n). An Olympus Vanta C series pXRF instrument containing a rhodium anode was used to measure XRF (0–40 keV) spectra in *Geochem* mode (2 beams) applying settings of 45 seconds for each beam and each sample was scanned duplicate (Fig. 2.4e).

6.2.4.3 Spectra pre-processing

Spectra collected from *in-situ* and laboratory scanning were pre-processed before further analysis. The data processing follows the method described in Chapters 2, 3 and 4. Briefly, from *in-situ* and laboratory vis-NIR spectra from both soil and plant samples, the noisiest part of the spectra 350–399 nm was discarded. In MIR spectra, only the 4000–605 cm^{-1} part of the spectra was used for predictive modelling. Reflectance spectra were derivated and smoothed using a Savitzky-Golay filter (first-order derivative, window size of nine and second-order polynomial) (Savitzky and Golay, 1964). From pXRF data, *beam 1* raw spectra were used after removing the first 24 wavelength data with zero spectral response values. The principal component analysis (PCA) of spectral data was performed to reduce the multi-dimensionality of spectral data into few principal components. Spectral data were centred before performing PCA. The distribution of experiments I and II soil, plant aboveground biomass, and root samples in the principal components space were plotted using the first two principal components describing the greatest amount of variance in the datasets.

6.2.5 Modelling framework

Predictive models to quantify Cd in soil (Pallic and Allophanic soils) and plant (aboveground biomass and root) samples were developed using the spectral information from both *in-situ* and laboratory scanning as input for PLS regression (Wold et al., 1983). Datasets used for predictive modelling included the reference laboratory Cd analysis results from experiment I Pallic soil (n=336) and aboveground biomass (n=28), and experiment II Allophanic soil (n=112), aboveground biomass (n=82), and root (n=28). During the model development, hyper-parametrisation of the prediction method tested was carried out using repeated k-fold cross-validation with 10 folds and 25 repeats, with

the optimal set of parameters chosen based on the simplest model within the one standard error of the empirically optimal model (Breiman et al., 1984).

6.2.6 Evaluation of model performance

The predictive accuracy of constructed models was assessed and compared using several performance statistics: RMSE, R^2, RPIQ (Bellon-Maurel et al., 2010), CCC (Lin, 1989), and bias. These parameters were determined using the R package *spectacles* (Roudier, 2021). The normalised RMSE value was calculated as a ratio of RMSE to mean Cd concentration (Pullanagari et al., 2016). Details on performance statistics can be found in Section 2.4.5 of the Chapter 2. The R package *caret* (Kuhn et al., 2021) was used to identify variable importance in projection contributing to predictions.

6.2.7 Calculations of vegetation indices, bioconcentration factors, and translocation factor

Seven vegetation indices (Table 6.2) relating to pigments (especially chlorophyll) and lignin content were calculated using *in-situ* reflectance spectra (Jordan, 1969). These indices were selected after assessing the correlation between wavelengths and plant Cd concentration.

The predicted Cd concentration for soil, aboveground biomass, and root samples were used to calculate bioconcentration and translocation factors. A prediction model using laboratory vis-NIR spectra as input for PLS was used to predict Pallic soil Cd. A model using laboratory pXRF spectra as input for PLS was used to predict Allophanic soil and root Cd. A model using *in-situ* spectra as input was used to predict plant aboveground biomass Cd concentration.

Table 6.2 Vegetation indices used in the study.

Vegetation indices	Formulas	Biological parameters	References
Blue green index 2	R_{450}/R_{550}	A low value indicates stress due to low chlorophyll or high lignin content.	(Zarco-Tejada et al., 2005)
Modified chlorophyll absorption ratio index 1	$((R_{700} - R_{670}) - 0.2 \times (R_{700} - R_{550})) \times (R_{700}/R_{670})$	A high value indicates low chlorophyll content, and high leaf area index.	(Daughtry et al., 2000)
Red green reflectance ratio	sum (R_{600}:R_{699}))/sum ((R_{500}:R_{599}))	A high value of this index indicates increase in anthocyanin pigment to detoxify accumulated Cd	(Dai et al., 2006; Gamon and Surfus, 1999)
Simple ratio index 6	$R_{672}/(R_{550} \times R_{708})$	A high value indicates high total carotenoids content maybe due to less Cd accumulation.	(Datt, 1998; Fan et al., 2023)
Transformed chlorophyll absorption in reflectance index	$3 \times ((R_{700} - R_{670}) - 0.2 \times ((R_{700} - R_{550}) \times (R_{700} / R_{670})))$	A high index value means less chlorophyll content.	(Haboudane et al., 2002)

Two Cd BCFs for each treatment and experiment were calculated as a ratio of chicory root or aboveground biomass to soil Cd concentration (Equations (2) and (3); Ait Ali et al. (2002)). The Cd TF for each treatment and experiment was calculated as a ratio of chicory aboveground biomass to root Cd concentration (Equation (4); Mattina et al. (2003)).

$$\text{Cd BCF root} = \frac{\text{Plant root Cd concentration (mg Cd/kg)}}{\text{Soil Cd concentration (mg Cd/kg)}} \quad (2)$$

$$\text{Cd BCF aboveground biomass} = \frac{\text{Plant aboveground biomass Cd concentration (mg Cd/kg DM)}}{\text{Soil Cd concentration (mg Cd/kg)}} \quad \text{.....(3)}$$

$$\text{Cd TF} = \frac{\text{Plant aboveground biomass Cd concentration (mg Cd/kg DM)}}{\text{Plant root Cd concentration (mg Cd/kg DM)}} \quad (4)$$

6.2.8 Statistical analyses

The R statistical environment was used to perform all statistical analyses (R Core Team, 2021). The distribution of the quantitative data was assessed using the Shapiro-Wilk normality test. Outliers were removed and non-normal data were transformed before further statistical analysis. Statistical summary of soil and plant aboveground biomass and root Cd concentrations in each experiment were determined and tabulated. The effect

of added soil Cd concentrations on the aboveground biomass of each harvest, total aboveground and root biomass in each experiment, plant Cd accumulation, bioconcentration and translocation factor values were analysed using one-way ANOVA analysis. For the significant ($p<0.05$) main effect, the difference between treatment means was tested using the least significance difference (LSD) test.

Correlation analysis between plant Cd concentration in the first harvest and *in-situ* spectra collected before the first harvest was performed using the *rcorr* function of the R package *Hmisc* (Harrell and DuPont, 2021). Regression analysis between vegetation indices and plant aboveground biomass Cd concentration, and measured and predicted TF, were performed and tabulated and/or presented graphically as needed.

6.3 Results

6.3.1 Plant biomass

One-way ANOVA showed no significant effect of amended soil Cd concentrations on the aboveground biomass in each harvest, the total aboveground and root biomass in each of the two experiments (Table 6.3). In experiment I, chicory average aboveground biomass was 3398 kg DM/ha, 1491 kg DM/ha, and 1545 kg DM/ha in the first, second, and third harvests, respectively (Table 6.3). The average total aboveground biomass was 6608 kg DM/ha, the average root biomass was 5952 kg DM/ha, and the average root: shoot ratio was 0.90 (Table 6.3). In experiment II, the average aboveground biomass was 486 kg DM/ha, 1259 kg DM/ha, and 1706 kg DM/ha in the first, second, and third harvests, respectively (Table 6.3). The average total aboveground biomass was 3415 kg DM/ha, the average root biomass was 3032 kg DM/ha, and the average root: shoot ratio was 0.87 (Table 6.3).

6.3.2 Soil and plant Cd concentrations

The background soil Cd concentration for the soils used in the study (0 mg added Cd/kg soil) was 0.21 mg Cd/kg for the Pallic soil (Experiment I) and 0.46 mg Cd/kg for the Allophanic soil (Experiment II) (Table 6.1). Amendment with Cd resulted in a

Table 6.3 Chicory temporal biomass in the glasshouse experiments I and II with amended increasing soil Cd concentrations.

Experiments	Added Cd concentrations (mg/kg soil)	First Harvest	Second Harvest	Third Harvest	Total aboveground biomass	Total root	First Harvest	Second Harvest	Third Harvest	Total aboveground biomass	Total Root	Root: Shoot ratio
		g DM/pot					kg DM/ha					
Experiment I	0	4.02	1.62	2.03	8.04	7.98	3196	1287	1609	6390	6343	0.99
	0.5	4.82	1.75	2.16	8.73	7.73	3828	1390	1718	6936	6137	0.88
	0.75	4.68	1.88	1.81	8.37	7.06	3714	1494	1440	6648	5607	0.84
	1.0	4.44	1.85	1.72	8.36	7.24	3528	1470	1367	6642	5748	0.87
	2.0	4.20	2.03	2.08	8.57	7.89	3335	1616	1651	6811	6265	0.92
	3.5	3.59	1.99	1.88	7.73	7.95	2850	1581	1492	6142	6317	1.03
	5.0	4.20	1.99	1.93	8.12	7.07	3337	1581	1535	6454	5619	0.87
	Average	4.28	1.88	1.94	8.32	7.49	3398	1491	1545	6608	5952	0.90
Added Cd concentration effect		ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
Experiment II	0	0.60	1.82	2.36	4.29	3.57	444	1340	1739	3158	2624	0.83
	0.5	0.69	1.61	1.94	4.23	4.04	506	1181	1424	3112	2974	0.96
	0.75	0.71	1.82	2.25	4.78	3.95	524	1336	1658	3518	2906	0.83
	1.0	0.73	1.62	2.22	4.57	3.99	537	1191	1632	3360	2939	0.87
	2.0	0.72	1.73	2.32	4.77	4.47	530	1273	1708	3511	3286	0.94
	3.5	0.64	1.68	2.46	4.78	4.52	467	1238	1811	3516	3323	0.95
	5.0	0.49	1.70	2.80	4.98	4.01	358	1249	2059	3666	2949	0.80
	Average	0.66	1.71	2.32	4.64	4.12	486	1259	1706	3415	3032	0.87
Added Cd concentration effect		ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns

ns=Non-significant

concentration range of 0.52 to 5.45 mg Cd/kg for experiment I and 0.84 to 4.81 mg Cd/kg soil for experiment II (Table 6.4). For chicory plants growing on the control treatment (0 mg added Cd/kg soil), the average Cd concentration was 3.93 mg Cd/kg DM in aboveground biomass for experiment I and 3.16 mg Cd/kg DM in aboveground biomass and 1.27 mg Cd/kg DM in root biomass for experiment II (Table 6.4; Appendix 6.2). Amendment of soil with increasing Cd concentrations resulted in an aboveground biomass Cd concentration range of 3.47 to 74.48 mg Cd/kg DM for experiment I and 2.40 to 56.77 mg Cd/kg DM for experiment II; and in a root biomass Cd concentration range of 0.86 to 25.79 mg Cd/kg DM for experiment II (Table 6.4; Appendix 6.2). Chicory aboveground biomass and root Cd concentration were significantly different among treatments ($p<0.05$) (Appendix 6.3).

Table 6.4 Measured soil and chicory aboveground and root biomass Cd concentrations in the glasshouse experiments I and II with amended increasing soil Cd concentrations.

Parameters	Experiment I			Experiment II		
	Soil	Chicory		Soil	Chicory	
		Aboveground biomass	*Root biomass		Aboveground biomass	Root biomass
	mg Cd/kg soil	mg Cd/kg DM		mg Cd/kg soil	mg Cd/kg DM	
Number of samples	336	28	–	112	82	28
Minimum	0.17	3.47	–	0.41	2.40	0.86
Mean	1.89	24.21	–	1.82	18.84	6.69
Maximum	5.45	74.48	–	4.81	56.77	25.79
Skewness	0.89	0.90	–	1.01	0.90	1.52

*No root samples were analysed from experiment I.

6.3.3 Cadmium bioconcentration and translocation factors

The Cd BCF and TF values for chicory were calculated using predicted Cd concentration of soil and chicory aboveground and root biomass (Fig. 6.3). The Cd BCF values for chicory root were significantly different among treatments in both experiments I ($p<0.01$) and II ($p<0.05$) (Fig. 6.3). The Cd BCF values were in the range of 2.6–12.4 (Figs 6.3a and 6.3b). In experiment I, BCF values were in the range of 2.8–12.4 (Fig. 6.3a). In experiment II, BCF values were in the range of 2.6–4.1 (Fig. 6.3b). The Cd BCF values

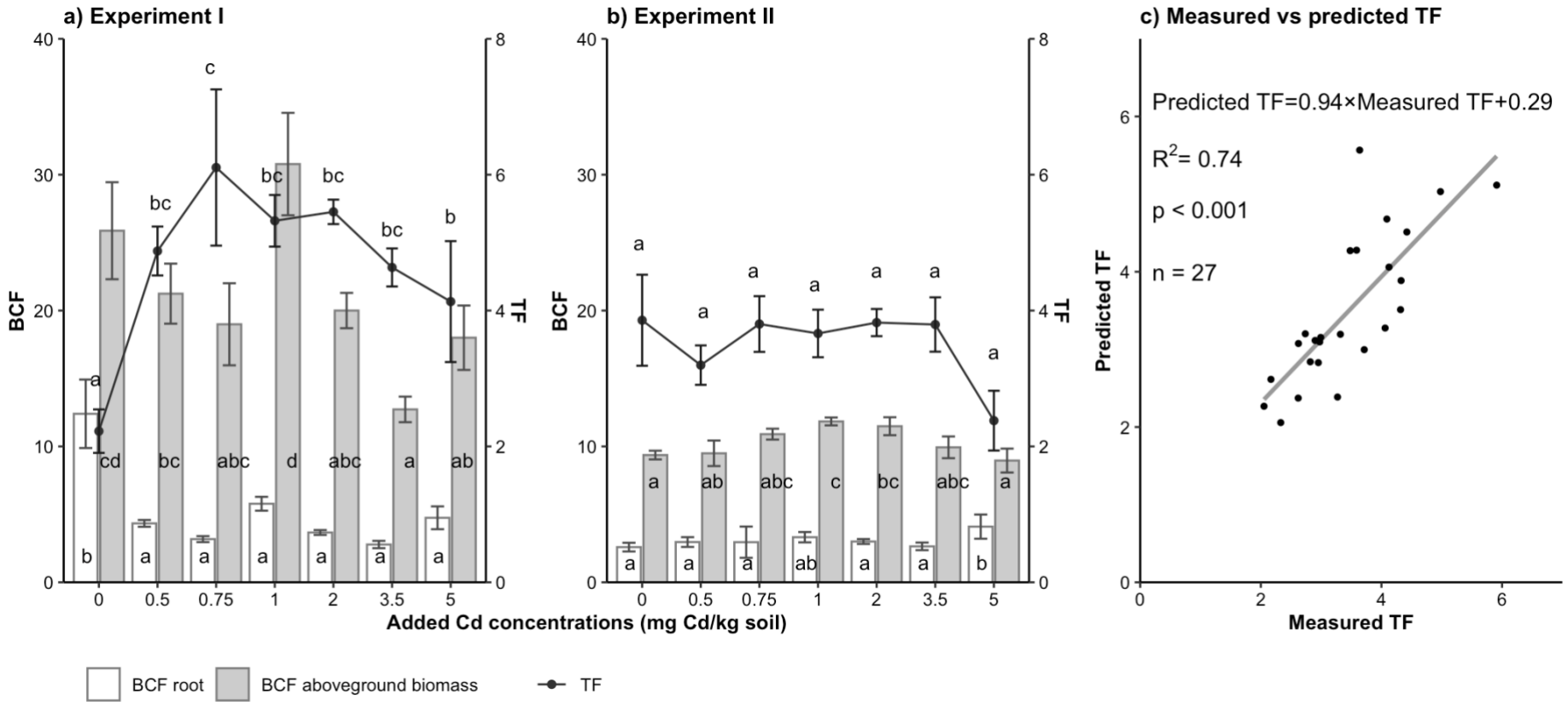


Fig. 6.3 Cadmium bioconcentration factors (BCF root, BCF aboveground biomass) and translocation factor (TF) values calculated using predicted Cd concentrations in the final soil and chicory total root and aboveground biomass from a) experiment I and b) experiment II. C) Regression analysis of predicted TF in relation to measured TF for chicory Cd accumulation in experiment II.

for chicory aboveground biomass values were significantly ($p<0.05$) different among treatments in both experiments (Figs 6.3a and 6.3b). The BCF values were in the range of 9.0–30.8 (Figs 6.3a and 6.3b). In experiment I, BCF values were in the range of 12.7–30.8 (Fig. 6.3a). In experiment II, BCF values were in the range of 9.0–11.8 (Fig. 6.3b).

The Cd TF values for chicory were in the range of 2.2–6.1 (Figs 6.3a and 6.3b). The TF values were significantly different among treatments in experiment I ($p<0.01$) (Fig. 6.3a). In experiment I, the TF values were in the range of 2.2–6.1 (Fig. 6.3a). In experiment II, TF values were in the range of 2.4–3.9 (Fig. 6.3b). A significant relationship ($R^2=0.74$, $p<0.001$, $n=27$) between predicted and measured TF was found for samples in experiment II (Fig. 6.3c).

6.3.4 Spectral similarity

The PCA plots presented in Figure 6.4 show the clustering of soil, aboveground biomass, and root samples in experiments I and II as per spectra characteristics (Fig. 6.4). Following PCA, the space defined by the first two principal components showed a clustering of soil and plant aboveground biomass samples into two clusters of experiment I and II separately as per laboratory vis-NIR, MIR, and pXRF spectra (Figs 6.4a and 6.4b), which can be interpreted as prediction models should be developed independently for each experiment. Whereas, following PCA, the space defined by the first two principal components showed some overlap between experiment I and II samples of soil and plant aboveground biomass as per *in-situ* vis-NIR spectra and of root biomass as per laboratory pXRF spectra (Figs 6.4a and 6.4b); this can be interpreted as a common prediction model can be developed combining spectral data from two experiments or a prediction model developed from one experiment can be used to predict the Cd concentration in the samples of another experiment. The first two principal components (PC1 and PC2) of soil pXRF spectra explained 99.5% spectral variance, with this value reduced to 98.5%, 85%, and 82% for the MIR, *in-situ* vis-NIR, and laboratory vis-NIR spectra, respectively (Fig. 6.4a). The first two principal components of aboveground biomass pXRF spectra explained 95% spectra variance, with this value reduced to 86%, 80%, and 67% for the laboratory vis-NIR, *in-situ* vis-NIR, and MIR spectra, respectively (Fig. 6.4b). The first two principal components of root pXRF spectra explained 96% spectral variance, with

this value reduced to 89% and 74% for the laboratory vis-NIR and MIR spectra, respectively (Fig. 6.4c).

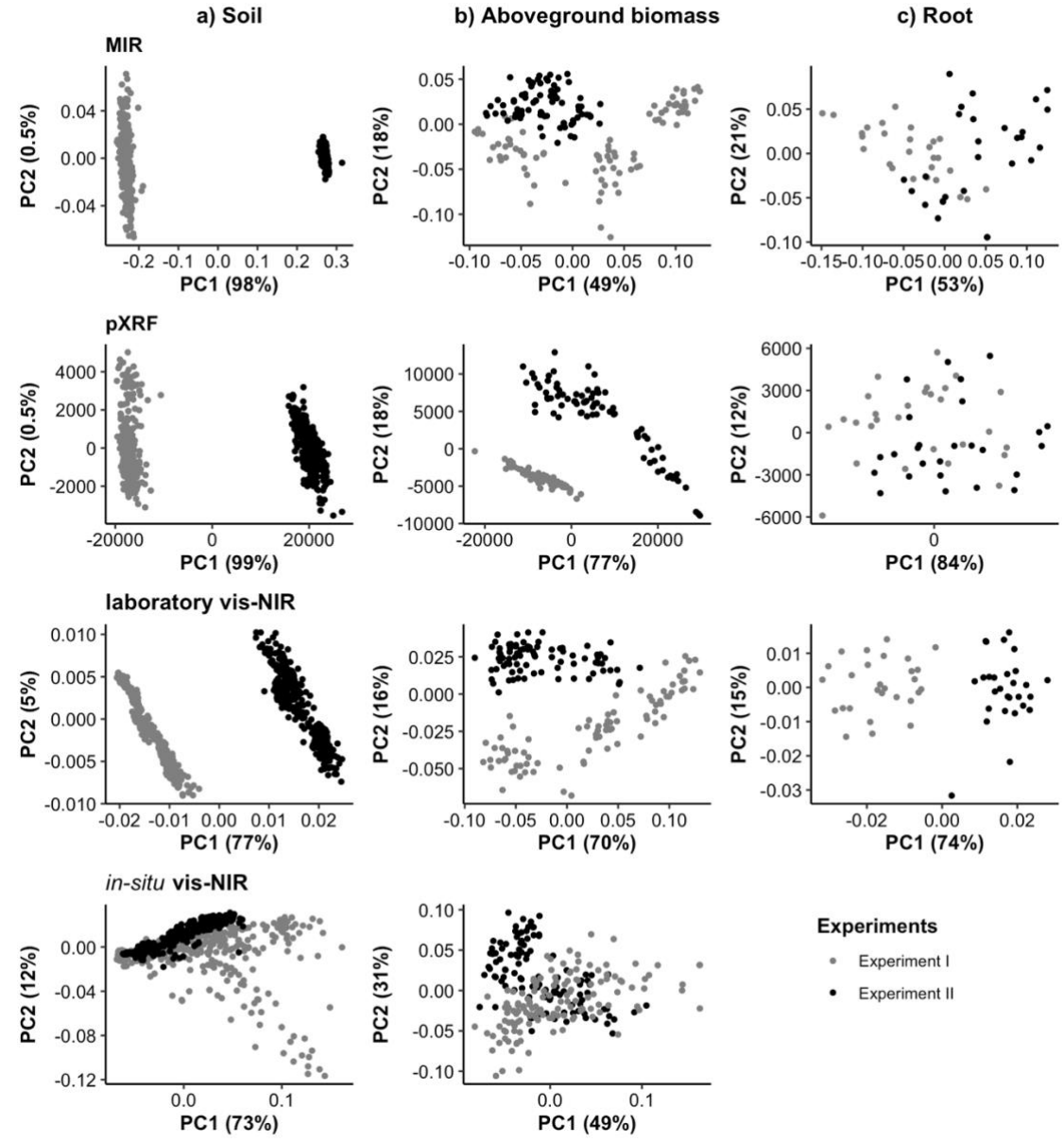


Fig. 6.4 Principal component analysis (PC1 and PC2) of laboratory and *in-situ* spectra for a) soil, b) aboveground biomass, and c) root samples from experiments I (Pallic soil: grey dots) and II (Allophanic soil: black dots).

6.3.5 Using *in-situ* spectra and vegetation indices to assess plant Cd concentration

Averaged *in-situ* spectra from aboveground chicory biomass showed variation in the spectral response in the 700–1350 nm region in both experiments (cell structure; Fig. 6.5a). In general, the reflectance of plants in this region showed a difference for those grown in both soils amended with more than 2 mg Cd/kg soil when compared to those with less than 2 mg Cd/kg (Fig. 6.5a). Linear correlations were significant (r=-0.8–0.7;

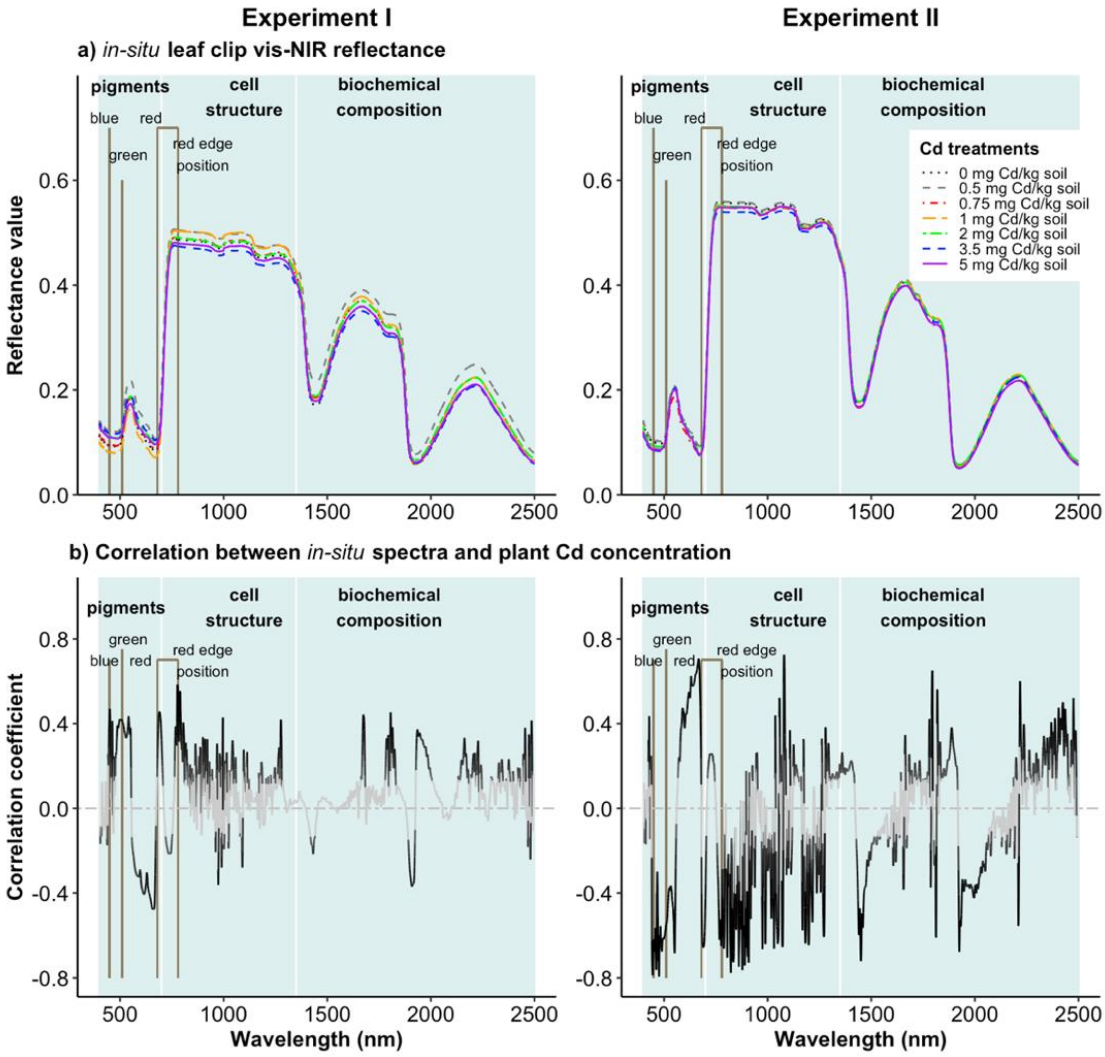


Fig. 6.5 a) *In-situ* vis-NIR reflectance spectra averaged for each Cd treatment, and b) linear correlations between transformed *in-situ* spectra wavelengths and plant aboveground Cd concentration measured before the first harvest in experiments I and II.

The black part of the line represents significant correlations ($p<0.05$) whereas the grey part represents non-significant correlations. The shaded area in the vis-NIR region: 350–700 nm (pigments), 700–1350 nm (cell structure), 1350–2500 nm (biochemical composition). Grey vertical lines represent 450 nm (blue), 510 nm (green), 680 nm (red), and 670–780 nm (red edge position).

$p<0.05$, n=28) between plant aboveground biomass Cd concentration in the first harvest and the specific regions of *in-situ* spectra collected before the first harvest in both experiments (Fig. 6.5b). Correlation values of more than 0.4 or less than -0.4 were significant in this analysis (Fig. 6.5b). In experiment I, a negative correlation ($r=-0.5$; $p<0.05$; n=28) was observed at 664 nm and a positive correlation ($r=0.6$; $p<0.05$; n=28) at 778 nm (Fig. 6.5b). In experiment II, a strong negative correlation ($r=-0.8$; $p<0.05$; n=28) (6.5b) was observed at 476 nm and a strong positive correlation ($r=0.7$; $p<0.05$; n=28) was observed at 1079 nm (Fig. 6.5b). A significant relationship ($R^2=0.15–0.36$; $p<0.05$) between vegetation indices and plant aboveground biomass Cd concentration was found in both experiments (Table 6.5). In experiment I, simple ratio index 6 showed the highest R^2 value of 0.23 (Table 6.5). In experiment II, blue green index 2 showed the highest R^2 value of 0.36 (Table 6.5).

Table 6.5 Regression analysis of vegetation indices in relation to aboveground biomass Cd concentration measured for the first harvest in experiments I (n=28; 3.47–74.48 mg Cd/kg DM) and II (n=27; 2.40–32.11 mg Cd/kg DM).

Vegetation indices	Experiment I		Experiment II	
	R^2	p-value	R^2	p-value
Blue green index 2	0.19	<0.05	0.36	<0.001
Modified chlorophyll absorption ratio index 1	0.18	<0.05	0.26	<0.01
Red green reflectance ratio	0.15	<0.05	0.30	<0.01
Simple ratio index 6	0.23	<0.01	0.25	<0.01
Transformed chlorophyll absorption in reflectance	0.20	<0.05	0.17	<0.05

6.3.6 Prediction of soil and plant Cd using laboratory spectra

Cross-validation prediction models were used to quantify Cd in soil and chicory aboveground biomass and root samples (Figs 6.6 and 6.7; Appendices 6.5, 6.6, 6.7, and

6.8). In general, models predicted well to quantify the Cd in root, while models quantifying Cd in soil and aboveground biomass performed variably (Figs 6.6 and 6.7; Appendices 6.5 and 6.6). Models using laboratory vis-NIR data performed satisfactorily quantifying Cd in soil and plant aboveground biomass and root (Figs 6.6 and 6.7; Appendices 6.5 and 6.6). Models developed using *in-situ* vis-NIR predicted plant aboveground biomass Cd well, whereas quantification was poor for soil Cd (Figs 6.6 and 6.7; Appendices 6.6). Models using MIR spectra performed consistently well quantifying Cd in all three matrices (Figs 6.6 and 6.7; Appendices 6.5 and 6.6). Models using pXRF data performed well to assess Cd even with a relatively low sample size (Allophanic soil: n=112; root: n=28) (Fig. 6.7; Appendix 6.6).

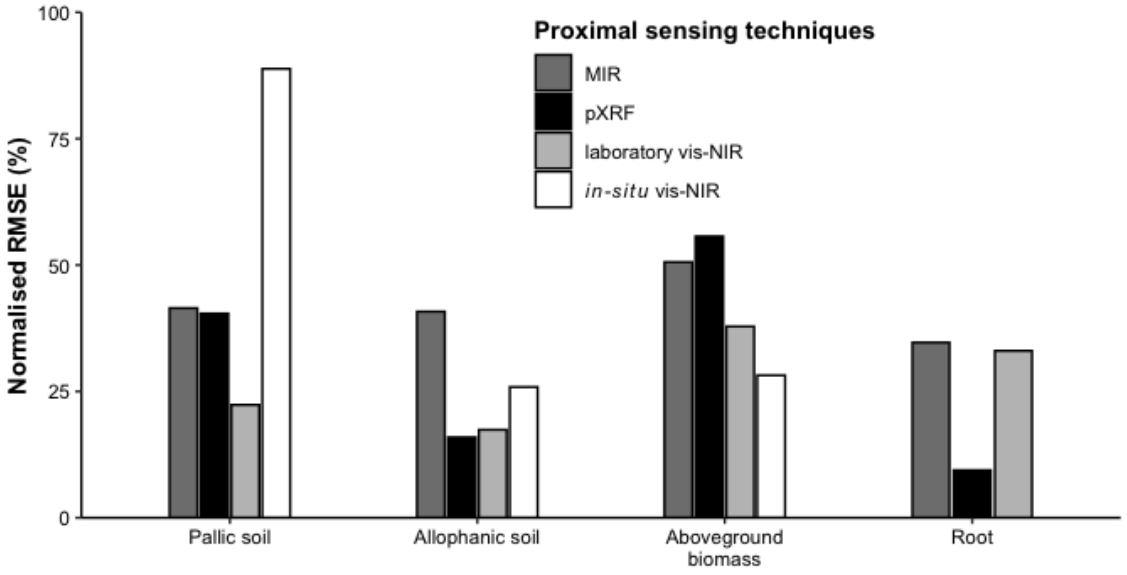


Fig. 6.6 Normalised root mean square error (RMSE) percentage for model prediction of total Cd in experiment I samples of soil (Pallic soil; n=224), and experiment II samples of soil (Allophanic soil; n=112), aboveground biomass (n=82), and root (n=28) using laboratory MIR, pXRF, and vis-NIR and *in-situ* vis-NIR sensor data as input for PLS. Except root samples, all other samples were also scanned *in-situ* using a vis-NIR sensor to develop Cd prediction models.

Models using laboratory vis-NIR spectra quantified Cd in Pallic soil well with $nRMSE_{cv}$ of 22% and CCC_{cv} of 0.97 (Figs 6.6 and 6.7; Appendix 6.5). Models using laboratory

pXRF and vis-NIR spectra quantifying Cd in Allophanic soil performed well (Figs 6.5 and 6.6; Appendix 6.6). These models showed $nRMSE_{cv}$ of 16% and 17% and CCC_{cv} of 0.95 and 0.97, respectively (Figs 6.5 and 6.6; Appendix 6.6). A model using *in-situ* vis-

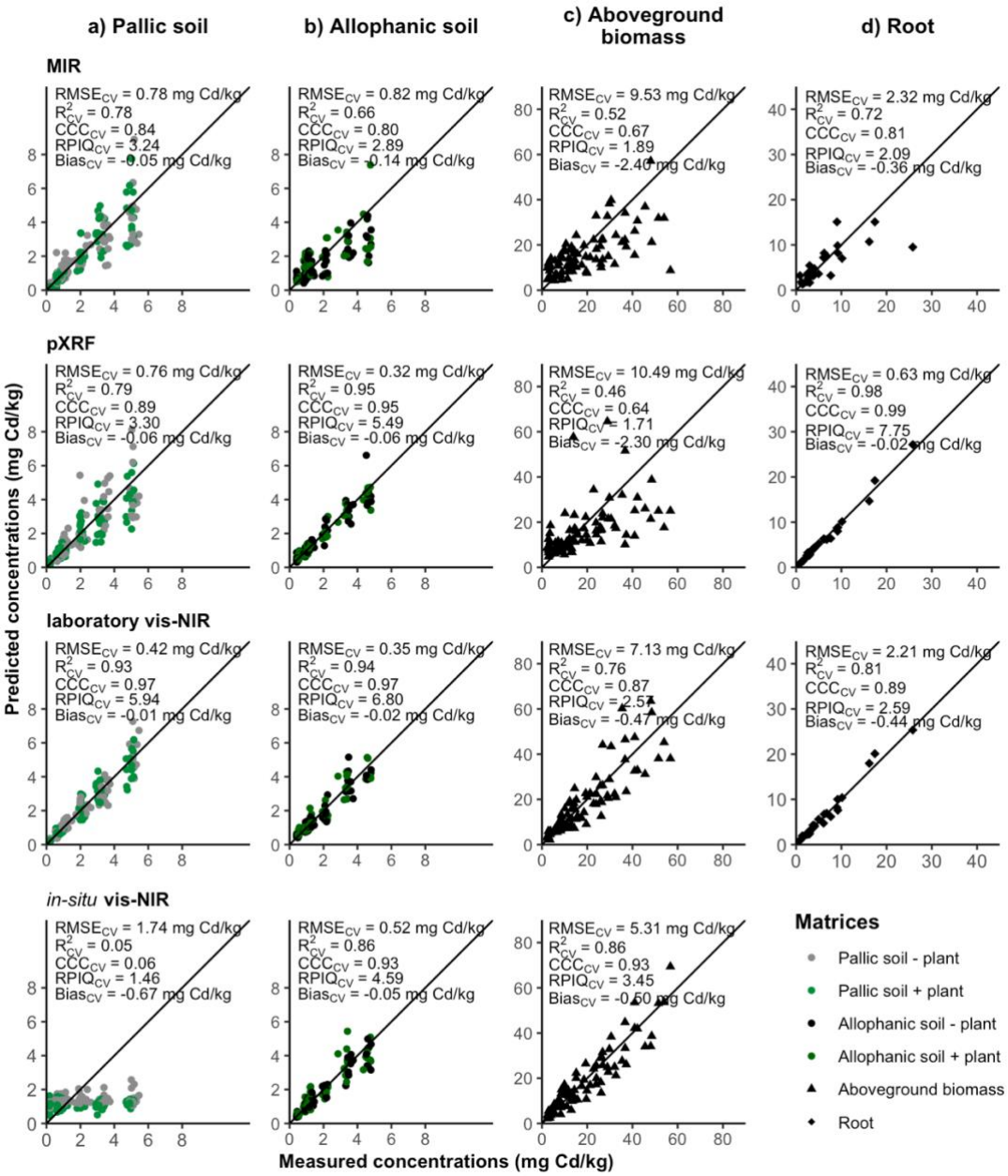


Fig. 6.7 Measured vs predicted concentration of Cd for the cross-validation set of a) experiment I samples of Pallic soil (n=224), and experiment II samples of b) Allophanic soil (n=112), and c) aboveground biomass (n=82) and root (n=28) using laboratory MIR, pXRF, and vis-NIR, and *in-situ* vis-NIR spectra as input for PLS.

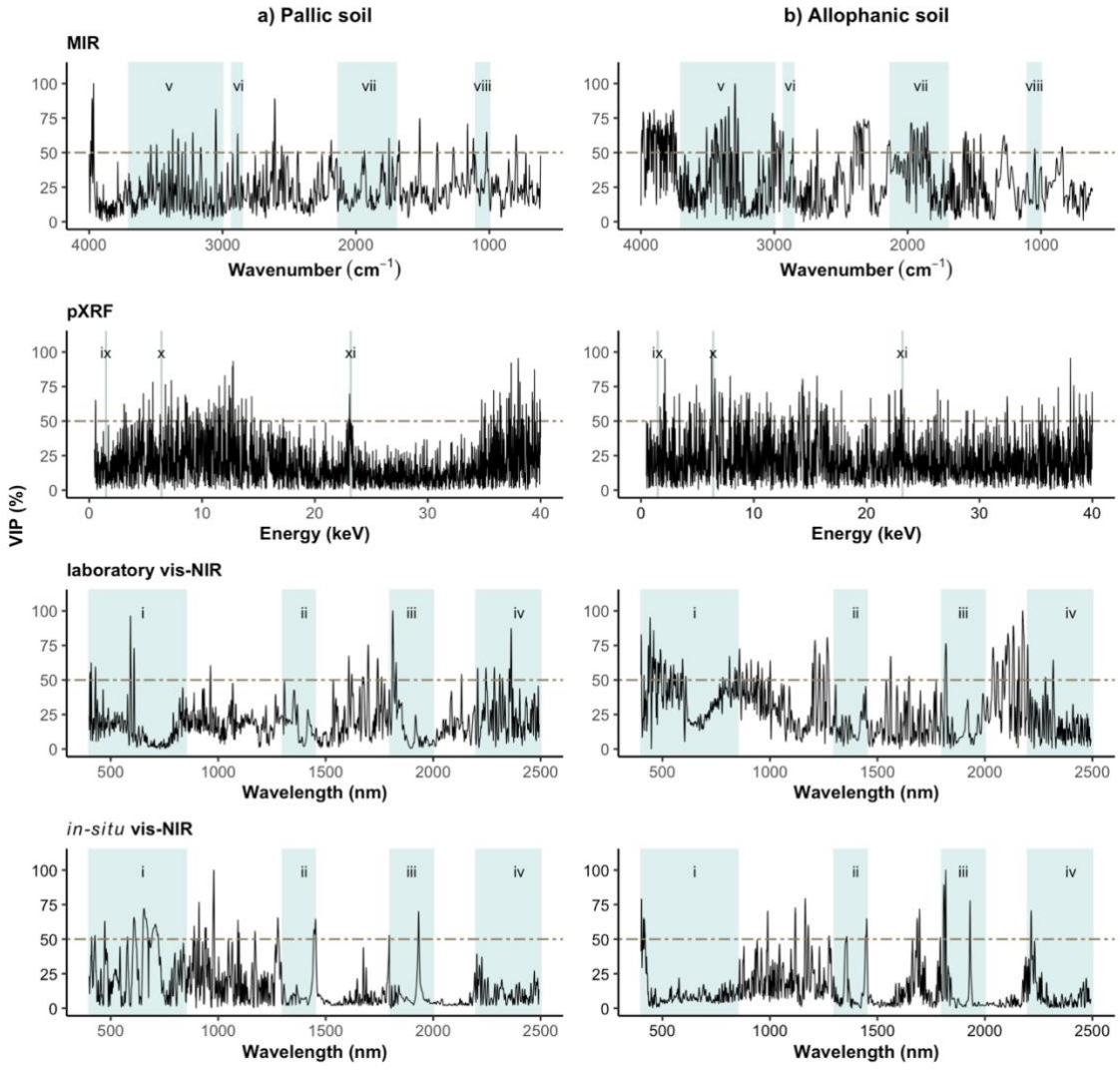


Fig. 6.8 Variable importance in projection (VIP) highlighting the importance of specific regions of laboratory MIR, pXRF, and vis-NIR, and *in-situ* vis-NIR spectra to quantify soil Cd in a) Pallic soil (experiment I) and b) Allophanic soil (experiment II). The shaded area in the vis-NIR region: i) soil colour, iron oxides, and soil organic matter (400−850 nm), ii) minerals (1300−1450 nm), iii) Al– and Fe–containing minerals and soil organic matter (1800−2000 nm), and iv) Al– and Fe–containing minerals and soil organic matter complex (2200−2500 nm); in the MIR region: v) Al– and Fe–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}); and in the XRF region: ix) Al (1.48 keV), x) Fe (6.40 keV), and xi) Cd (23.17 keV).

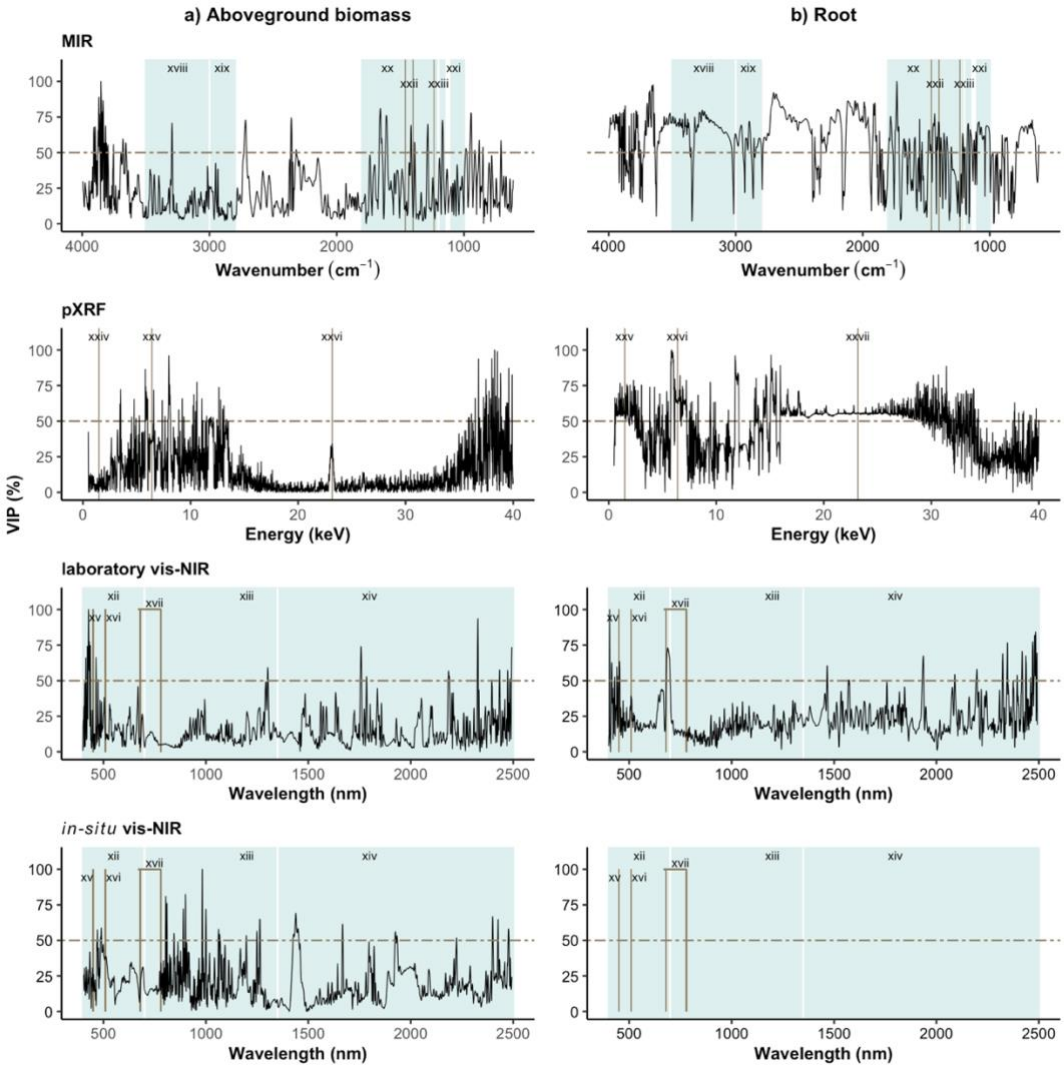


Fig. 6.9 Variable importance in projection (VIP) highlighting the importance of specific regions of laboratory MIR, pXRF, and vis-NIR, and *in-situ* vis-NIR spectra to predict plant Cd in a) aboveground biomass and b) root. The shaded area in the vis-NIR region: xii) pigments (350–700 nm), xiii) cell structure (700–1350 nm), xiv) biochemical composition (1350–2500 nm). Grey vertical lines represent xv) blue (450 nm), xvi) green (510 nm), xvii) red (680 nm) and the red edge position (670–780 nm). In the MIR region: xviii) O–H and N–H stretching vibration (3500–3000 cm^{-1}), xix) CH_3 and CH_2 stretching vibration (3000–2800 cm^{-1}), xx) ester-containing compounds commonly found in membrane lipid and cell wall pectin; amide I and II in proteins and vibrations of aromatic ring like lignin derivatives (1800–1200 cm^{-1}), xxi) cellulose in the leaves characterised by C–H bending or C–O or C–C stretching (1100–1000 cm^{-1}). Grey vertical lines represent xxii) CH_3 and CH_2 bending motion (1460–1400 cm^{-1}), xxiii) C–O stretching in ester and amide III (1235–1153 cm^{-1}). In the XRF region: xxv) Al (1.48 keV), xxvi) Fe (6.40 keV) and xxvii) Cd (23.17 keV).

NIR spectra predicting Cd in plant aboveground biomass performed reasonably well ($nRMSE_{cv}$=28%, CCC_{cv}=0.93; Figs 6.5 and 6.6; Appendix 6.8). A prediction model using pXRF data predicting Cd in root performed very well with $nRMSE_{cv}$ of 9% and CCC_{cv} of 0.99 (Figs 6.5 and 6.6; Appendix 6.4).

The PLS variable importance in projection highlighted the relative importance of certain spectral regions, and these areas contributed to explain the performance of a particular prediction model (Figs 6.8 and 6.9). To quantify Cd in Pallic soil, variable importance in projection showed the importance of several regions of laboratory vis-NIR spectra: 400–850 nm (iron oxides and soil organic matter), 1800–2000 nm (Al– and Fe–containing minerals and soil organic matter), and 2200–2500 nm (Al–OH or Fe–OH bearing minerals and soil organic matter complex) regions (Fig. 6.8a; Appendix 2.1). To quantify Cd in aboveground biomass, variable importance in projection showed the importance of the regions 700–1350 nm (cell structure) and 1350–2500 nm (biochemical composition) in *in-situ* vis-NIR spectra (Fig. 6.9a). To quantify Cd in Allophanic soils and root, variable importance in projection showed the importance of the Al (1.48 keV), Fe (6.40 keV), and Cd (23.17 keV) regions in pXRF spectra (Fig. 6.8b and 6.9b).

6.4 Discussion

In this Chapter, three proximal sensors (vis-NIR, MIR, and pXRF) were assessed independently to quantify Cd in agricultural soil and chicory (aboveground biomass and root) samples (Figs 6.6 and 6.7). This study evaluated the effect of wide concentrations of soil Cd on plant biomass and Cd uptake, applied vegetation indices as an indicator of plant Cd concentration, and quantified soil and plant Cd using *in-situ* and laboratory scanning data.

6.4.1 Effect of amended soil Cd in plant biomass, plant accumulation, and factors of Cd phyto-availability

Plant biomass was not affected by the added increasing Cd concentrations in the current study (Table 6.3), which is in agreement with Ubeynarayana et al. (2021), who found a non-significant difference in chicory aboveground biomass after 60 days (0.25–0.33 g DM/pot) for a soil Cd concentrations of up to 1.6 mg Cd/kg. Stafford et al. (2016) also

showed no significant effect of plant Cd concentration (1.48–1.96 mg/kg DM) on aboveground biomass (1511–1775 kg DM/ha). In addition, coefficient of determination (R^2) values obtained between vegetation indices and plant aboveground biomass Cd concentration in the current study were low when compared to those described in studies using toxic Cd concentrations (Feng et al., 2019; Götze et al., 2010; Wang et al., 2018). Feng et al. (2019) determined R^2 value of 0.70 between a vegetation index and aboveground biomass Cd concentration in a range of 1 to 1046 mg Cd/kg DW. Kooistra et al. (2004) calculated R^2 value of 0.49 between a vegetation index and soil samples taken from grass vegetation (n=18) with a mean Cd concentration of 7.3 mg Cd/kg soil. To avoid the toxic effect of less than 10 mg Cd/kg soil concentration, plant accumulated Cd is generally complexed with organic acids and transported to the cell wall, vacuoles, or leaf trichomes (Lai and Chen, 2006; Sharma and Dubey, 2006). Deposition of Cd in cell structure, leaf thickness, or intercellular space was indicated by variable importance in projection for aboveground biomass, and a significant correlation between Cd concentration and 750–1350 nm region of *in-situ* leaf clip vis-NIR spectra (Figs 6.3 and 6.8; Bandaru et al. (2010); Gitelson et al. (2006); Gitelson et al. (2001)).

Accumulation of Cd in chicory (aboveground biomass and root) samples was high in this study, particularly when compared to those values found in the previous studies (Table 6.5; Anderson et al. (2022); Stafford et al. (2016); Yang et al. (2021)). For a soil amended with Cd concentration of 1.6 mg Cd/kg soil, Ubeynarayana et al. (2021) found chicory accumulated a Cd concentration of 19 mg Cd/kg in aboveground biomass and 7 mg Cd/kg in root biomass after 60 days of plant growth. In another glasshouse experiment, the Cd concentration in chicory was 12 mg Cd/kg aboveground biomass after 49 days of plant growth in a soil Cd concentration of 1.21 mg Cd/kg soil (Crush et al., 2019). In contrast, in a field survey, Anderson et al. (2022) found chicory plants growing on soil with 0.40 mg Cd/kg accumulated a Cd concentration of less than 1 mg Cd/kg DM. In a pot experiment where Cd was added through phosphate fertiliser, Stafford et al. (2016) found at 0.46 mg Cd/kg soil, chicory accumulated 1.86 mg Cd/kg DM. A field survey by Cavanagh et al. (2019) showed the Cd concentration in onion shoot grown on non-Allophanic soil (0.011 mg/kg fresh weight) was higher in amount than Allophanic soil (0.015 mg/kg fresh weight) when compared to their soil Cd concentration (0.14 mg Cd/kg fresh soil and 0.77 mg Cd/kg fresh soil, respectively). The reason for the higher uptake and accumulation of Cd in chicory in the current study could be due to the presence of a

higher phyto-available concentration of Cd in the rhizosphere soil (Ubeynarayana et al., 2021).

Soil pH, organic matter content, CEC, and soil type could have influenced Cd phyto-availability. McBride and Li (2022) conducted a soybean bioassay in contaminated soil (total Cd: 36.7 mg Cd/kg soil and modified Morgan extractable Cd: 7.97 mg Cd/kg soil) from an orchard where sewage sludge was applied 40 years before and found Cd phyto-availability was still high even after 40 years of application due to reducing soil pH (6.26 to 5.97) and declined soil organic matter content (12.3% to 8.72%). In a 6-year long field experiment, Jiang et al. (2018) found that phyto-availability of trace elements (Cu, Ni) increased in soil at pH 5.3. Marković et al. (2019) used a shaking method (10 revolutions per minute) to extract Cd from the soil at a solution ratio of 1 g to 20 ml over two months and found that fractional distribution was significantly correlated to soil pH, total organic carbon, and CEC of the soil. At the relatively low soil pH in the soils used in this study (Experiment I, Pallic soil, soil pH=5.30; Experiment II, Allophanic soil, soil pH=5.42), there is increased activity of acidic cations (H^+, Fe^{3+}, Al^{3+}) in the soil solution, which compete with Cd^{2+} for pH-dependent sites (e.g., electrostatic negatively charged sites of clay and organic matter surfaces contributing to CEC), promoting Cd phyto-availability (Table 6.1; Appendix 6.1; Cottenie and Verloo (1984)).

6.4.2 Bioconcentration and translocation factor values

This study calculated bioconcentration and translocation factors for chicory using proximal sensor data, as laboratory vis-NIR data for soil, *in-situ* vis-NIR data for chicory aboveground biomass and pXRF data for chicory root biomass based on PCA analysis and/or prediction model results (Figs 6.4 and 6.7). The significant R^2 value was obtained between measured and predicted TF values in this study (Fig. 6.3), which suggests the potential of using vis-NIR (for aboveground and soil) and pXRF (root) sensor data in evaluating Cd food chain transfer risk as well as Cd mass balance models (Gray and Cavanagh, 2022). Both BCF and TF values calculated in this study were within the range of the previous studies (Fig. 6.9; Table 6.6; Abe et al. (2008); Guérin et al. (2022); Stafford et al. (2016); Yang et al. (2021)). In the previous studies, for plant species including chicory, perennial ryegrass, and Amaranthus grown in Cd concentrations of 0.2–11 mg Cd/kg soil, Cd BCF for root were in the range of 0.9 to 1011, Cd BCF for

aboveground biomass were in the range of 0.1 to 453, and Cd translocation factor were in the range of 0.3 to 29 (Table 6.6; Abe et al. (2008); Guérin et al. (2022); Stafford et al. (2016); Yang et al. (2021)). In all studies, both BCF and TF values were influenced by phyto-availability of Cd (added) in soil, the total soil Cd, soil properties, type, and duration of the experiment (i.e., glasshouse or field experiment), status of essential soil nutrients, and plant characteristics (Table 6.6). Plant's defence mechanism drives the flow of Cd uptake from the soil in the root to aboveground biomass to dilute the toxic effect of Cd, which can be indicated by the TF values (Lai and Chen, 2006).

Table 6.6 Cadmium bioconcentration factor (BCF) values for root and aboveground biomass, and translocation factor (TF) values for chicory determined in the selected previous studies.

Soil Cd concentration (mg Cd/kg)	Duration of the experiment (days)	BCF root	BCF aboveground	TF	References
0.17	60	127	453	3.6	(Abe et al., 2008)
0.46	266	–	4.0	–	(Stafford et al., 2016).
1.21	49	–	7.4–10.9	–	(Crush et al., 2019)
1.60	60	–	–	4.4	(Ubeynarayana et al., 2021)
5.09	224	0.9	0.3	3.2	(Yang et al., 2021)
10.7	156	2.3–2.9	0.3–0.4	6.3–28.8	(Guérin et al., 2022)

6.4.3 *In-situ* scanning to assess plant Cd

The deployment of *in-situ* leaf clip scanning could ease the rapid monitoring of Cd in pastoral farms. The accuracy achieved by models to predict aboveground biomass Cd concentration using *in-situ* vis-NIR leaf clip spectra in the current study (Fig. 6.7) is comparable to previous studies that have used plant canopy, airborne imaging, and laboratory hyperspectral imaging (Feng et al., 2019; Kooistra et al., 2004; Pullanagari et al., 2016). Kooistra et al. (2004) developed a PLS model using canopy vis-NIR spectra and predicted Cd concentration in grass samples with R^2_{cv} value of 0.52 and $nRMSE_{cv}$ of 18%. Pullanagari et al. (2016) used hyperspectral airborne imaging data (380–2500 nm) as input for PLS and predicted trace elements concentrations in the mixed pasture with R^2_{cv} value of 0.68 and $nRMSE_{cv}$ of 17%. Feng et al. (2019) used vis-NIR hyperspectral imaging data as input for modified PLS and predicted leaf Cd with R^2_{v} value of 0.91 and

$nRMSE_v$ of 20%. Compared to other *in-situ* spectra capturing methods, the leaf clip method is comparatively easy and fast, and obtained spectra contain detailed and specific information on pigment content, cell structure, and biochemical composition (Gholizadeh and Kopackova, 2019; Rathod et al., 2013). In this study, changes in cell structure which could be related to the deposition of Cd also suggested by (1) variable importance in projection showed the significance of the 700–1350 nm region of *in-situ* vis-NIR spectra for aboveground biomass Cd prediction (Fig. 6.9; Kooistra et al. (2004)), (2) a significant correlation between the 700–1350 nm region of the vis-NIR spectra and plant aboveground Cd concentration might be associated with the interaction of Cd with cell wall polysaccharides and cell wall plasticity (Fig. 6.3b; Sharma and Dubey (2006)) and (3) reduced reflectance at higher Cd concentration which may be due to reduced leaf thickness and intercellular space causing decreased multiple scattering of incident 700–1350 nm rays (Fig. 6.3a; Bandaru et al. (2010); Fu et al. (2020); Rosso et al. (2005)).

6.4.4 Vegetation index as a rapid Cd detection tool

This study tested previously developed vegetation indices to rapidly detect Cd in plant aboveground biomass (Tables 6.2 and 6.5). Among vegetation indices tested, "blue green index 2" developed by Zarco-Tejada et al. (2005), using a ratio of reflectance values at 450 and 550 nm wavebands (Table 6.2), showed potential to use as a rapid tool to detect Cd in chicory (Table 6.5). Zarco-Tejada et al. (2005) found a significant relationship between blue green index 2 and the sum of chlorophyll a and chlorophyll b content in leaf. An increase in Cd content has been shown to reduce chlorophyll a and chlorophyll b pigment content in plant leaf (Fan et al., 2023). Scanning plants at an early stage of growth (day 60 in experiment II) resulted in a more significant relationship between index and plant Cd than at later stages (day 73 in experiment I) (Table 6.5; Gamon and Surfus (1999)). Additionally, the results of this work suggest that greater plant biomass in experiment I has diluted the Cd accumulation effect relative to experiment II (Table 6.3; Fan et al. (2023)).

6.4.5 Laboratory scanning to assess soil and plant Cd

In the literature, laboratory studies can be found comparing the three proximal sensors used in the current study to assess soil properties, including Cd concentration (Chapters

3, 4, and 5; O'Rourke et al. (2016a); Shrestha et al. (2022)). In addition to assess soil Cd, this study evaluated these three proximal sensors to quantify Cd in plant samples (Figs 6.6 and 6.7). The performance of prediction models developed using pXRF data to quantify Cd in the root is in agreement with McGladdery et al. (2018), who used pXRF spectra data as input for simple linear regression to develop a Cd prediction model to accurately quantify Cd concentration (10–265 mg Cd/kg DM) in powdered plant samples with R^2_v value of 0.75 and $RMSE_v$ of 0.07 mg Cd/kg. A reasonably good quantification of Cd in root samples may be related to (1) Cd concentration in the root matrix to be detected by the fluorescence energy of the sensor and (2) a low concentrations of interfering elements (e.g., Ca, K, Fe, Zn) in the root matrix (Elam et al., 2002; Nwafor et al., 2017; Stosnach, 2005).

Differences in the performance of prediction models developed to quantify Cd in two soil types using one of three proximal sensors can be related to the difference in: (1) the association of soil Cd with spectrally active soil components involved in its retention, i.e., soil organic matter and Al–, Fe–, and Mn–containing minerals (Demattê et al., 2004; Kooistra et al., 2001), as indicated by relevant PLS variable importance in projection of vis-NIR spectra for Cd in Pallic soil (Fig. 6.7a), whereas in case of Allophanic soil pXRF spectral response to the soil Cd concentration and Al–, and Fe–containing minerals (Figs 6.7b and 6.8b; Kalnicky and Singhvi (2001); O'Rourke et al. (2016a)).

6.5 Conclusions

This study successfully assessed a wide range of soil and plant Cd concentrations by analysing reflectance and fluorescence spectra. Although, amended soil Cd concentrations showed no significant effect in chicory biomass, *in-situ* leaf clip vis-NIR spectra showed differences in the 700–1350 nm (cell structure) region. There was also a significant correlation between spectral bands in the 700–1350 nm region of vis-NIR spectra and plant Cd concentrations. The blue green index 2, a vegetation index, can potentially be used for rapid Cd assessment on pastoral farms. This study suggests that vis-NIR and pXRF sensors can be used to evaluate food chain transfer risks associated with Cd accumulation in edible plant parts. Prediction models developed using proximal sensor data in this study could potentially contribute to cost-effective Cd monitoring in pastoral farms.

Chapter 7

Key findings of the study, their implications for New Zealand agriculture, and opportunities for future work

7.1 Introduction

Implementation of proximal sensing techniques can potentially be a cost-effective and rapid alternative to monitor Cd levels for effective Cd management in agricultural soils according to the tiered fertiliser management system. Although application of such techniques has shown success in monitoring elevated Cd concentrations in mining and contaminated areas, these techniques are understudied to quantify low Cd concentration in agricultural soils (Song et al., 2012; Stafford et al., 2018b). Spectral data from vis-NIR, MIR, and pXRF has been successfully used to quantify Cd concentrations at local and national scales (O'Rourke et al., 2016b; Siebielec et al., 2004). However, the possibility of using global prediction models for Cd prediction at a local-scale is poorly explored (Ng et al., 2022a). In addition to quantifying Cd concentration in soil, using proximal sensing techniques to assess Cd levels in forage plant species such as chicory, may improve the understanding of Cd food chain transfer risks from agricultural systems (Gray and Cavanagh, 2022).

This doctoral study was conducted to develop robust prediction models using proximal sensing techniques including vis-NIR, MIR, and pXRF sensors independently or in combination, allowing accurate assessment of low Cd concentration from regional- to farm-scale. In this Chapter, key findings from the research work performed over a scale range from the region, farm, and pot and with a wide range of Cd concentrations are integrated and discussed in the context of the knowledge gap described in the Chapter 2. To enhance the predictive power of proximal sensing techniques in this work,

chemometric methods including sensors data fusion, model fusion, model averaging, and memory-based learning algorithms were used.

7.2 Key findings

Combinations of proximal sensors (pXRF, vis-NIR, and/or MIR) data and chemometric techniques successfully predicted low concentrations of eight TE in regional scale samples (Chapter 3)

Topsoil (0–20 cm) samples (n=622) obtained from a previous geochemical baseline survey from sites spaced 8 km apart covering c. 40,000 km^2, were accurately quantified for low concentrations of eight soil TE (As, Cd, Cr, Cu, Hg, Pb, Ni, and Zn) by prediction models using combinations of pXRF, vis-NIR, and/or MIR sensors data. A chemometric method PLS-SVM model fusion that has been successfully applied to assess other soil properties including available water-holding capacity and soil organic matter content (Bao et al., 2017; Blaschek et al., 2019), was applied along with data fusion or model averaging techniques to build prediction models using this large dataset that performed with better accuracy than previously published studies (Table 3.3). The current research has shown the PLS-SVM can accurately predict TE using proximal sensors, particularly when applying data fusion and/or model averaging (Fig. 3.3; Appendix 3.2). The PLS algorithm performs matrix decomposition on the spectral and TE concentration data to extract latent variables and this helps to reduce data redundancy and remove noise and matrix interference caused mainly by soil organic matter in XRF spectra (Bao et al., 2017; Blaschek et al., 2019; Hong et al., 2019; Ravansari and Lemke, 2018). The use of SVM regression on PLS latent variables considers the non-linear spectral association between TE and spectrally active soil components, contributing to successful predictions (Blaschek et al., 2019).

The results of reference laboratory analysis of TE concentrations were compared against the built-in pXRF calibration model. Built-in pXRF calibration performed with comparable accuracy for Pb (Fig. 3.3; Appendices 3.2 and 3.5). For other TE (As, Cd, Cr, Cu, Hg, Ni, and Zn), recalibration of the pXRF with local samples was required to yield accurate results (Fig. 3.3; Appendices 3.2 and 3.4). Models using spectral data fusion of pXRF with vis-NIR or MIR reflectance data as input for PLS-SVM outperformed models

using pXRF data to quantify As, Cr, Pb, Cu, Ni, and Zn (Fig. 3.3; Appendices 3.2 and 3.4). Models using pXRF spectra as input performed reasonably with some deviations from the measured concentrations due to the spectral interference caused by soil matrix components including organic matter and other mineral components such as Mg, Ca, Al, Fe, and Mn (Appendices 3.2 and 3.4; Elam et al. (2002); Stosnach (2005)). Both organic matter and minerals contribute to reflectance spectra and regulate TE binding in soils, revealing indirectly associated TE; and thus TE quantification was more accurate and precise when reflectance spectra were combined with pXRF spectra (Fig. 3.4; Appendices 3.2 and 3.4; Kooistra et al. (2001); Wu et al. (2010)).

Granger-Ramanathan model averaging improved the estimation of Cd and Hg, assigning a higher combination weight to the vis-NIR model output (Fig. 3.3; Appendix 3.4). This is explained by PLS loadings which highlight the association of Cd and Hg with soil organic matter and Al containing minerals, and the association of Cd with Fe oxides and Al– and Fe– containing minerals in vis-NIR spectra (Fig. 3.4c; Appendices 2.1 and 3.4; Kemper and Sommer (2002)).

Each of the proximal sensing techniques vis-NIR, MIR, and pXRF showed comparable accuracy in assessing total soil Cd, reflectance spectroscopy techniques showed the potential to predict plant-available and potentially-available Cd fractions in agricultural soils (Chapter 4)

Topsoil (0−15 cm) samples (n=87) from 30 pastoral farms with long-term phosphate fertiliser application history from across the North and South Islands were analysed for total Cd and the distribution of Cd among geochemical fractions (Fig. 4.1; Table 2.3). The main soil components affecting Cd accumulation were Al containing mineral surfaces and soil organic matter (Table 4.2; Loganathan et al. (2012); Yuan and Lavkulich (1997)). Cation exchange capacity strongly influenced the mobility of Cd sorbed to soil organic matter (Table 4.2; Amacher et al. (1986); Stafford et al. (2018c)). However, despite soil pH being a major determining factor of CEC, in this study, soil pH had no significant influence on total soil Cd concentration within the topsoil samples (Table 4.2; Salmanzadeh et al. (2016)).

For this farm scale dataset, PLS-based cross-validation (which excludes the need to maintain separate calibration and validation sets) prediction models were developed using

individual sensor data as input (Chodak et al., 2007). The prediction model developed in this work quantifying total soil Cd based on vis-NIR data relied on the association of Cd with spectrally active soil components: (1) soil organic matter and (2) Al– and Fe–containing minerals (Kooistra et al., 2001; Song et al., 2012; Xu et al., 2020) as indicated by (1) PLS variable importance in projection (VIP) of the selected regions including spectral response from organic matter and minerals (Fig. 4.8a) and (2) significant correlations between total soil Cd and total C, total N, acid oxalate extractable Al and Fe (Table 4.2; Gholizadeh et al. (2021); Christensen and Haung (1999); Young (2013)). Models quantifying total Cd based on MIR data relied on the association of Cd with Al– and Fe–containing minerals and metal–carbonyl bonds in the matrix (Janik et al., 1998; Wang et al., 2017), and this association is supported by variable importance in projection (VIP) of the selected regions including spectral response from Al– and Fe–containing minerals and metal–carbonyl bonds (Fig. 4.8a). Models quantifying total Cd based on pXRF relied on the direct relationship between pXRF spectral response and Cd concentration (Chapter 3; O'Rourke et al. (2016a)).

Reflectance spectroscopy showed potential feasibility as a technique to predict plant-available and potentially-available Cd fractions in agricultural soils. The accuracy of a Cd prediction model using vis-NIR data to assess acid soluble and organic matter bound Cd fractions (potentially-available fractions for plant uptake) relied on the association of these fractions with soil organic matter (Gray et al., 2000; Kooistra et al., 2001; Krishnan et al., 1980) as indicated by (1) PLS variable importance in projection (VIP) of selected regions including spectral response from soil organic matter (Figs 4.8c and 4.8e) and (2) significant correlations between total soil Cd and total C and total N (Table 4.2). The predictive accuracy achieved by a model quantifying exchangeable Cd using MIR spectra was influenced by (1) association between exchangeable Cd and Al– and Fe–containing minerals and/or soil organic matter (Fig. 4.8b; Stenberg and Rossel (2010); Wang et al. (2017)), (2) water bridging mechanism (Fig.4.8b; Chakraborty et al. (2017); Sposito (1984)), and (3) the possible presence of exchangeable Cd in inorganic forms (e.g., $CdCl^{+}$, $Cd(NO_3)_2$) which does not influence reflectance spectra (Bolan et al., 1999; Li et al., 2021; Moros et al., 2009). Metal oxides bound and residual Cd fractions were not accurately predicted (Figs 4.7d and 4.7f; Table 4.3; Chakraborty et al. (2017)) because of (1) occurrence of the significantly higher amount of reactive Al, Fe, and Si in Allophanic

soil but the limited correlation of Cd with these mineral components (Figs 4.5b, 4.6e, and 4.6f) and (2) very low concentration range of residual Cd (Fig. 4.7f; Xie et al. (2012)).

A large-scale soil spectral library quantified soil Cd concentration in local agricultural soil samples (Chapter 5)

To test the possibility of using a SSL to quantify total Cd in local soil samples, a SSL developed from the Otago and Southland regional survey (n=625; 0.01–1.31 mg Cd/kg) data and the local set containing agricultural soils from a pastoral farm survey (n=87; 0.10–2.03 mg Cd/kg; Chapter 4) were used. Individual sensor data-based prediction models developed using PLS and LOCAL algorithms were customised to predict local Cd concentration (agricultural samples) while leveraging a SSL. Spiking the regional SSL pastoral soil subset with selected local samples (particularly in combination with selection and weighing: RP283+L12×4; Figs 5.6 and 5.7) provided an additional advantage when predicting Cd locally, improving accuracy (reduced RMSE) by 42% compared to only SSL subset based prediction models (Fig. 5.6; Appendices 5.2 and 5.3). The improvement in Cd prediction was due to (1) spiking the large SSL pastural soils subset similar to the local set with 12 local samples including Allophanic and Pumice soils that had a known Cd concentration of ≥0.6 mg Cd/kg, thus filling key gaps in the database (pedodiversity and concentration) (Brown, 2007; Guerrero et al., 2010; Wetterlind and Stenberg, 2010) and (2) similarity of Cd association with Al– and Fe–containing minerals (3700–3000 cm^{-1}) and metal–carbonyl bonds (2100–1730 cm^{-1}) in both sets as observed in the PLS loadings (Fig. 5.7c; Janik et al. (1998); Niazi et al. (2015); Wang et al. (2017)).

In general, noise from overtones captured by the vis-NIR spectra from the fundamental vibrations of molecular bonds and functional groups (Fig. 5.7a; Siebielec et al. (2004); Stenberg and Rossel (2010)) limited the accuracy of models using this sensor data (Figs 5.6b and 5.6f; Appendix 5.2). The poor performance of models using pXRF data (Figs 5.6d and 5.6h) to quantify Cd in local samples may be mainly related to the low Cd concentration range of the regional SSL samples (average 0.08 mg. Cd/kg soil; Fig. 5.1b) (Lemière, 2018; Weindorf and Chakraborty, 2020).

Significant R^2 value was obtained between measured and predicted Cd translocation factor values for chicory using proximal sensor data (Chapter 6)

Two independent glasshouse experiments were conducted using Pallic and Allophanic soils amended with increasing Cd concentrations and with or without a model forage herb, chicory (*Cichorium intybus* L.). Chicory biomass was not affected by the added Cd concentrations in these experiments (Table 6.3; Stafford et al. (2016); Ubeynarayana et al. (2021)). This study calculated translocation factor (TF) values using proximal sensor data (Figs 6.3a and 6.3b) and showed a significant R^2 value between measured and predicted TF values (Fig. 6.3c).

For this dataset, cross-validation prediction models were developed using individual sensor data as input for PLS algorithm. A prediction model developed using *in-situ* leaf clip vis-NIR spectra showed optimal performance to assess aboveground biomass Cd concentration (Fig. 6.7). The accuracy of this technique was related to reflectance spectra capturing changes in cell structure and leaf thickness which could be related to cellular or intercellular deposition of Cd as indicated by (1) variable importance in projection highlighted the importance of the 700–1350 nm region of *in-situ* leaf vis-NIR spectra which is related to cell structure (Fig. 6.9; Kooistra et al. (2004)), (2) significant correlation between the 700–1350 nm region of the vis-NIR spectra and plant aboveground Cd concentration which might be associated with the interaction of Cd with cell wall polysaccharides and cell wall plasticity (Fig. 6.3b; Sharma and Dubey (2006)), and (3) reduced reflectance at higher Cd concentrations which may be due to reduced leaf thickness and intercellular space causing decreased multiple scattering of incident 700–1350 nm rays (Fig. 6.3a; Bandaru et al. (2010); Fu et al. (2020); Rosso et al. (2005)).

The R^2 value obtained between vegetation indices and plant aboveground biomass Cd concentration in this study was low when considered against the range of values (R^2=0.15–0.36; Table 6.5) recorded in studies using toxic Cd concentrations (R^2=0.49–0.70; Feng et al. (2019); Kooistra et al. (2004)). Among the vegetation indices calculated, "blue green index 2" developed by Zarco-Tejada et al. (2005) showed a significant and higher R^2 value (0.36, for experiment II) for plant aboveground Cd concentration than the other vegetation indices determined (Table 6.5).

The performance of prediction models developed in the current study using pXRF data to quantify the Cd concentration in the root (Figs 6.6 and 6.7), was in agreement with

McGladdery et al. (2018), who used pXRF spectra data as the input for simple linear regression to develop a Cd prediction model to accurately quantify Cd concentration (10–265 mg Cd/kg DM) in powdered plant samples with R^2_v value of 0.75 and $RMSE_v$ of 0.07 mg Cd/kg. A reasonably good quantification of Cd in the root samples may be related to a low concentrations of interfering elements (e.g., Ca, K, Fe, Zn) in the root matrix (Elam et al., 2002; Nwafor et al., 2017; Stosnach, 2005).

Differences in the performance of prediction models developed to quantify Cd in two soil types using one of three proximal sensors can be related to differences in the association of soil Cd with spectrally active soil components involved in its retention, i.e., soil organic matter, and Al–, Fe–, and Mn–containing minerals (Bolan et al., 2003a; Demattê et al., 2004; Kooistra et al., 2001), as indicated by relevant PLS variable importance in projection (VIP) of vis-NIR spectra for Cd in Pallic soil (Fig. 6.7a) and in case of Allophanic soil, the spectral response of pXRF to soil Cd concentration and Al– and Fe–containing minerals (Figs 6.7b and 6.8b; O'Rourke et al. (2016a); Kalnicky and Singhvi (2001)).

7.3 Implications for Cd monitoring and management in New Zealand agriculture

This doctoral study has shown that proximal sensing techniques can rapidly analyse Cd concentration in soil and plant samples (Chapters 3, 4, and 6). Optimal prediction models were developed in this study used proximal sensor data independently or in combination, and these models can be deployed for long-term Cd monitoring at local-to-national scales (Chapters 3 and 4). The level of precision and accuracy reported in this study makes Granger-Ramanathan model averaging approach a good candidate that could feasibly be deployed for long-term monitoring of soil Cd at concentration below the pXRF detection limits and with reduced matrix interference from organic matter when compared to the individual techniques alone (Fig. 3.3; Appendix 3.2). The systematic combination approach taken in this study could be further used to (1) reduce the economic costs of sample preparation and analytics, (2) facilitate detailed monitoring of Cd concentration on individual farms for sustainable soil management, as well as (3) extend monitoring beyond total Cd concentration to include an assessment of Cd fractions associated with

soil components such as soil organic matter and Al– and Fe–containing minerals (Chapter 3).

The use of individual proximal sensing techniques (vis-NIR, MIR, and pXRF) to determine total soil Cd in New Zealand agricultural soils has the potential to improve the scale and scope of long-term repeated monitoring of soil Cd required under the framework of the TFMS (Figs 4.6 and 4.7a; Table 4.3). The reflectance spectroscopy techniques (vis-NIR, and/or MIR) showed a good technical feasibility in their ability to predict Cd fractions in agricultural soils and this could be implemented to monitor soil Cd fractions for minimising plant Cd uptake (Figs 6.6 and 6.7). Prediction models using vis-NIR data assessing acid soluble and organic matter bound Cd are relevant to the study of plant Cd accumulation as both are potentially-available geochemical forms for plant uptake (Chavez et al., 2016; El-Mufleh Al Husseini et al., 2013; Gray et al., 2000). Models predicting exchangeable Cd using MIR could be feasibly deployed to classify agricultural soils with respect to the tier-based management system (Fig. 4.7b; Cécillon et al. (2009)) (Chapter 4).

The optimal prediction model using MIR data from the regional SSL pastoral soil subset (RP283) spiked with weighed local samples quantifying Cd in local samples (Fig. 5.6e; Appendix 5.3) could be implemented for environmental monitoring, such as the TFMS designed to monitor and manage Cd concentration in New Zealand agricultural soils (Gray and Cavanagh, 2022). Such approach could reduce the analytical cost to the farmers and allow intensive spatial and temporal monitoring of pastoral farms based on spectral analysis only (Shepherd and Walsh, 2002). In the current study, the optimal model included 12 out of 87 (14%) local samples analysed using reference laboratory method, while remaining samples were assessed using proximal sensing techniques. Given that a large SSL is available, increase in the analytical cost can be considered moderate (Chapter 5).

Proximal sensor data, as *in-situ* leaf clip vis-NIR and laboratory pXRF, can be used to estimate Cd bioconcentration and translocation factor values in plants (Fig. 6.3). The significant R^2 value between measured and predicted TF values obtained in this study shows the potential of using proximal sensor data to evaluate food chain transfer risks (Fig. 6.3c). A vegetation index, blue green index 2, can potentially be used as a rapid Cd detection tool for chicory at early stage of growth (Chapter 6).

The spectral library developed from the study containing soil and plant root and aboveground biomass pXRF, vis-NIR, and MIR spectra with a wide range of Cd concentration (Chapters 3, 4, and 6) can be used as reference ground data for airborne to satellite-based monitoring (Nawar and Mouazen, 2017b; Pullanagari et al., 2016; Shepherd et al., 2022).

7.4 Opportunities for future work

This doctoral study allows the definition of key opportunities for the wider implementation of proximal sensing techniques for cost-effective analysis of Cd and other properties based on the reported success in developing robust prediction models using proximal sensor data including vis-NIR, MIR, and pXRF independently and in combination for accurate assessment of a wide range of Cd concentration found in New Zealand agricultural soil and plant samples. Optimal prediction models developed from this study could be successfully employed for soil and plant Cd assessment in the laboratory (or other controlled environments) using vis-NIR, MIR, and pXRF sensors. There is also an opportunity to use these proximal sensors to quantify total Cd on the farm without requiring samples to be taken to the laboratory. However, field-level validation of the optimal models developed in the current study would be first needed, considering the influence of soil moisture, radiation, particle size, and other environmental, plant, and edaphic factors. The Otago-Southland spectral library was developed from a regular 8 km spaced soil samples collected at the regional scale, and there is a good possibility of using this SSL for local scale Cd monitoring using the regular spacing sampling protocol of the survey. The transferability of the developed spectral library should be examined by collecting samples in regular spacing at a pastoral farm to prepare a Cd distribution map of the target farm. In addition, extending the potential of proximal sensors to assess Cd geochemical fractions by including more samples covering soil types, land uses, and geographic locations could potentially contribute to an effective system to monitor Cd fractions in agricultural soils to minimise plant Cd uptake and subsequent food chain transfer risks. For rapid plant Cd assessment, there is a good potential to apply the vegetation index “blue green index 2” in vegetable crops, but again this needs to be tested and verified. Cadmium uptake by vegetable crops and accumulation in edible plant parts directly influence human health. Developing proximal sensor-based prediction models to quantify Cd concentration in market gardens soils and their vegetable produces could

contribute to reduce health and trade risks. Furthermore, the transferability of Cd prediction models using *in-situ* vis-NIR leaf clip spectra to canopy spectra and then to airborne hyperspectral imaging has the potential to contribute to large-scale monitoring of Cd across agricultural farms.

There is considerable interest in deploying the proximal sensors researched in this thesis to precision agriculture, soil and plant physical, chemical, and biological analysis, and the real-time monitoring of crop status, land use, and carbon by airborne and remote sensing platforms (Shepherd et al., 2022; Whitehead et al., 2021). The realisation of opportunities for future work suggested by this doctoral study could potentially contribute to developing economically and environmentally sustainable agricultural systems in New Zealand without losing the land use flexibility of highly productive lands that define the contribution of agriculture to New Zealand's economic prosperity.

References

Abbott, D., 2014. Applied predictive analytics: Principles and techniques for the professional data analyst. A John Wiley & Sons Ltd Publication.

Abe, T., Fukami, M., Ogasawara, M., 2008. Cadmium accumulation in the shoots and roots of 93 weed species. Soil Sci. Plant Nutr. 54(4), 566-573. https://doi.org/10.1111/j.1747-0765.2008.00288.x.

Abraham, E., 2018. Cadmium in New Zealand agricultural soils. N. Z. J. Agric. Res. 63(2), 202-219. https://doi.org/10.1080/00288233.2018.1547320.

Acharjee, A., 2012. Comparison of regularized regression methods for ~Omics data. J. Postgenomic Drug Biomarker Dev. 03(03). https://doi.org/10.4172/2153-0769.1000126.

Adriano, D., 1986. Trace elements in the terrestrial environment. . New York, Springer Verlag.

Adriano, D.C., 1992. Biogeochemistry of trace metals: Advances in trace substances research. CRC Press.

Ait Ali, N., Bernal, M.P., Ater, M., 2002. Tolerance and bioaccumulation of copper in Phragmites australis and Zea mays. Plant Soil 239(1), 103-111. https://doi.org/10.1023/A:1014995321560.

Alloway, B.J., 2013. Heavy metals in soils. Environ. Pollut., 22. Springer Science & Business Media. https://doi.org/10.1007/978-94-007-4470-7.

Amacher, M.C., Kotuby-Amacher, J., Selim, H.M., Iskandar, I.K., 1986. Retention and release of metals by soils — evaluation of several models. Geoderma 38(1), 131-154. https://doi.org/10.1016/0016-7061(86)90011-X.

Anderson, C.W.N., Smith, S.L., Jeyakumar, P., Thompson-Morrison, H., Cavanagh, J.-A.E., 2022. Forage crops and cadmium: How changing farming systems might impact cadmium accumulation in animals. Sci. Total Environ., 154256. https://doi.org/10.1016/j.scitotenv.2022.154256.

Angelopoulou, T., Dimitrakos, A., Terzopoulou, E., Zalidis, G., Theocharis, J., Stafilov, T., Zouboulis, A., 2017. Reflectance spectroscopy (Vis-NIR) for assessing soil heavy metals concentrations determined by two different analytical protocols, based on ISO 11466 and ISO 14869-1. Water Air Soil Pollut. 228(11), 436. https://doi.org/10.1007/s11270-017-3609-9.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J.-L., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J. Hydrol. 529, 754-767. https://doi.org/10.1016/j.jhydrol.2015.09.001.

Artz, R.R., Chapman, S.J., Robertson, A.J., Potts, J.M., Laggoun-Défarge, F., Gogo, S., Comont, L., Disnar, J.-R., Francez, A.-J., 2008. FTIR spectroscopy can be used as a screening tool for organic matter quality in regenerating cutover peatlands. Soil Bio. Biochem. 40(2), 515-527. https://doi.org/10.1016/j.soilbio.2007.09.019.

Asgari, N., Ayoubi, S., Jafari, A., Demattê, J.A.M., 2020. Incorporating environmental variables, remote and proximal sensing data for digital soil mapping of USDA soil great groups. Int. J. Remote Sens. 41(19), 7624-7648. https://doi.org/10.1080/01431161.2020.1763506.

ATSDR, 2012. Toxicological profile for cadmium, Agency for Toxic Substances and Disease Registry, Public Health Service, US Department of Health and Human Services.

Bacon, J.R., Davidson, C.M., 2008. Is there a future for sequential chemical extraction? Analyst 133(1), 25-46. https://doi.org/10.1039/b711896a.

Baldock, J.A., Hawke, B., Sanderman, J., Macdonald, L.M., 2013. Predicting contents of carbon and its component fractions in Australian soils from diffuse reflectance mid-infrared spectra. Soil Res. 51(7-8), 577-583. https://doi.org/10.1071/SR13077.

Bandaru, V., Hansen, D.J., Codling, E.E., Daughtry, C.S., White-Hansen, S., Green, C.E., 2010. Quantifying arsenic-induced morphological changes in spinach leaves: Implications for remote sensing. Int. J. Remote Sens. 31(15), 4163-4177. https://doi.org/10.1080/01431161.2010.498453.

Bao, N., Wu, L., Ye, B., Yang, K., Zhou, W., 2017. Assessing soil organic matter of reclaimed soil from a large surface coal mine using a field spectroradiometer in laboratory. Geoderma 288, 47-55. https://doi.org/10.1016/j.geoderma.2016.10.033.

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. Trac-Trend Anal. Chem. 29(9), 1073-1081. https://doi.org/10.1016/j.trac.2010.05.006.

Ben-Dor, E., 2002. Quantitative remote sensing of soil properties. Adv. Agron. 75, 173-243. https://doi.org/10.1016/S0065-2113(02)75005-0.

Ben-Dor, E., Banin, A., 1990a. Diffuse reflectance spectra of smectite minerals in the near infrared and their relationship to chemical composition. Sci. Geol. Bull. 43(2-4), 117-128. https://doi.org/10.3406/sgeol.1990.1848.
Ben-Dor, E., Banin, A., 1990b. Near-infrared reflectance analysis of carbonate concentration in soils. Appl. Spectrosc. 44(6), 1064-1069. https://doi.org/10.1366/0003702904086821.
Ben-Dor, E., Banin, A., 1995. Near infrared analysis (NIRA) as a method to simultaneously evaluate spectral featureless constituents in soils. Soil Sci. 159(4), 259-270.
Benedet, L., Faria, W.M., Silva, S.H.G., Mancini, M., Dematte, J.A.M., Guilherme, L.R.G., Curi, N., 2020. Soil texture prediction using portable X-ray fluorescence spectrometry and visible near-infrared diffuse reflectance spectroscopy. Geoderma 376, 114553. https://doi.org/10.1016/j.geoderma.2020.114553.
Blackburn, G.A., 1998. Quantifying chlorophylls and caroteniods at leaf and canopy scales: An evaluation of some hyperspectral approaches. Remote Sens. Environ. 66(3), 273-285. https://doi.org/10.1016/S0034-4257(98)00059-5.
Blakemore, L.C., Searle, P.L., Daly, B.K., 1987. Methods for chemical analysis of soils. NZ Soil Bureau scientific report: 80. Rev. ed. NZ Soil Bureau, Department of Scientific and Industrial Research.
Blaschek, M., Roudier, P., Poggio, M., Hedley, C.B., 2019. Prediction of soil available water-holding capacity from visible near-infrared reflectance spectra. Sci. Rep. 9(1), 12833. https://doi.org/10.1038/s41598-019-49226-6.
Bolan, N.S., Adriano, D.C., Mani, P.A., Duraisamy, A., 2003a. Immobilization and phytoavailability of cadmium in variable charge soils. II. Effect of lime addition. Plant Soil 251(2), 187-198. https://doi.org/10.1023/a:1023037706905.
Bolan, N.S., Adriano, D.C., Naidu, R., 2003b. Role of phosphorus in (im)mobilization and bioavailability of heavy metals in the soil-plant system, Reviews of Environmental Contamination and Toxicology: Continuation of Residue Reviews. Springer New York, pp. 1-44. https://doi.org/10.1007/0-387-21725-8_1.
Bolan, N.S., Khan, M.A.R., Tillman, R.W., Naidu, R., Syers, J.K., 1999. The effects of anion sorption on sorption and leaching of cadmium. Aust. J. Soil Res. 37(3), 445-460. https://doi.org/10.1071/s97046.
Bray, J.G.P., Rossel, R.V., McBratney, A.B., 2009. Diagnostic screening of urban soil contaminants using diffuse reflectance spectroscopy. Aust. J. Soil Res. 47(4), 433-442. https://doi.org/10.1071/Sr08068.
Breiman, L., 2001. Random forests. Mach. Learn. 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. Routledge.
Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A., Jordan, A., 2016. Soil mapping, classification, and pedologic modeling: History and future directions. Geoderma 264, 256-274. https://doi.org/10.1016/j.geoderma.2015.05.017.
Brown, C.D., Vega-Montoto, L., Wentzell, P.D., 2000. Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration. Appl. Spectrosc. 54(7), 1055-1068. https://doi.org/10.1366/0003702001950571.
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(4), 444-453. https://doi.org/10.1016/j.geoderma.2007.04.021.
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.
Brownlee, J., 2019. A tour of machine learning algorithms.
Buschmann, C., Nagel, E., 1993. In vivo spectroscopy and internal optics of leaves as basis for remote sensing of vegetation. Int. J. Remote Sens. 14(4), 711-722. https://doi.org/10.1080/01431169308904370.
Campos, M.P., Reis, M.S., 2020. Data preprocessing for multiblock modelling – A systematization with new methods. Chemometr. Intell. Lab. Sys. 199, 103959. https://doi.org/10.1016/j.chemolab.2020.103959.
Cao, Y., Sun, J., Yao, K., Xu, M., Tang, N., Zhou, X., 2021. Nondestructive detection of lead content in oilseed rape leaves based on MRF-HHO-SVR and hyperspectral technology. J. Food Process Eng. 44(9), e13793. https://doi.org/0.1111/jfpe.13793.
Caporale, A.G., Adamo, P., Capozzi, F., Langella, G., Terribile, F., Vingiani, S., 2018. Monitoring metal pollution in soils using portable-XRF and conventional laboratory-based techniques: Evaluation

of the performance and limitations according to metal properties and sources. Sci. Total Environ. 643, 516-526. https://doi.org/10.1016/j.scitotenv.2018.06.178.
Cariati, F., 1983. Polarization of water molecules in phyllosilicates in relation to exchange cations as studied by near infrared spectroscopy. Clay Clay Miner. 31(2), 155-157. https://doi.org/10.1346/ccmn.1983.0310211.
Cavanagh, J., 2014. Status of cadmium in New Zealand soils, Landcare Research, Lincoln.
Cavanagh, J., McNeill, S., Arienti, C., Rattenbury, M., 2015. Background soil concentrations of selected trace elements and organic contaminants in New Zealand.
Cavanagh, J.E., Yi, Z., Gray, C.W., Munir, K., Lehto, N., Robinson, B.H., 2019. Cadmium uptake by onions, lettuce and spinach in New Zealand: Implications for management to meet regulatory limits. Sci. Total Environ. 668, 780-789. https://doi.org/10.1016/j.scitotenv.2019.03.010.
CCME, 1999. Canadian soil quality guidelines for the protection of enviornmental and human health: Cadmium, Canadian environmental quality guidelines, 1999. Canadian Council of Ministers of the Environment, Winnipeg, Canada.
Cécillon, L., Barthès, B.G., Gomez, C., Ertlen, D., Genot, V., Hedde, M., Stevens, A., Brun, J.J., 2009. Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). Eur. J. Soil Sci. 60(5), 770-784. https://doi.org/10.1111/j.1365-2389.2009.01178.x.
Chakraborty, R., Kereszturi, G., Pullanagari, R., Durance, P., Ashraf, S., Anderson, C., 2022. Mineral prospecting from biogeochemical and geological information using hyperspectral remote sensing - Feasibility and challenges. J. Geochem. Explor. 232, 106900. https://doi.org/10.1016/j.gexplo.2021.106900.
Chakraborty, S., Li, B., Deb, S., Paul, S., Weindorf, D.C., Das, B.S., 2017. Predicting soil arsenic pools by visible near infrared diffuse reflectance spectroscopy. Geoderma 296, 30-37. https://doi.org/10.1016/j.geoderma.2017.02.015.
Chakraborty, S., Weindorf, D.C., Li, B., Ali Aldabaa, A.A., Ghosh, R.K., Paul, S., Nasim Ali, M., 2015. Development of a hybrid proximal sensing method for rapid identification of petroleum contaminated soils. Sci. Total Environ. 514, 399-408. https://doi.org/10.1016/j.scitotenv.2015.01.087.
Chang, C.W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65(2), 480-490. https://doi.org/10.2136/sssaj2001.652480x.
Chappelle, E.W., Kim, M.S., McMurtrey, J.E., 1992. Ratio analysis of reflectance spectra (RARS): An algorithm for the remote estimation of the concentrations of chlorophyll A, chlorophyll B, and carotenoids in soybean leaves. Remote Sens. Environ. 39(3), 239-247. https://doi.org/10.1016/0034-4257(92)90089-3.
Chavez, E., He, Z.L., Stoffella, P.J., Mylavarapu, R.S., Li, Y.C., Baligar, V.C., 2016. Chemical speciation of cadmium: An approach to evaluate plant-available cadmium in Ecuadorian soils under cacao production. Chemosphere 150, 57-62. https://doi.org/10.1016/j.chemosphere.2016.02.013.
Chen, J.M., 1996. Evaluation of vegetation indices and a modified simple ratio for boreal applications. Can. J. Remote Sens. 22(3), 229-242. https://doi.org/10.1080/07038992.1996.10855178.
Chen, L., Larson, S.L., Ballard, J.H., Ma, Y., Zhang, Q., Li, J., Wu, L., Arslan, Z., Han, F.X., 2019. Laboratory spiking process of soil with various uranium and other heavy metals. 6, 734-739. https://doi.org/10.1016/j.mex.2019.03.026.
Chen, T., Chang, Q., Clevers, J.G., Kooistra, L., 2015. Rapid identification of soil cadmium pollution risk at regional scale based on visible and near-infrared spectroscopy. Environ. Pollut. 206, 217-226. https://doi.org/0.1016/j.envpol.2015.07.009.
Cheng, H., Shen, R.L., Chen, Y.Y., Wan, Q.J., Shi, T.Z., Wang, J.J., Wan, Y., Hong, Y.S., Li, X.C., 2019. Estimating heavy metal concentrations in suburban soils with reflectance spectroscopy. Geoderma 336, 59-67. https://doi.org/10.1016/j.geoderma.2018.08.010.
Chodak, M., Niklinska, M., Beese, F., 2007. Near-infrared spectroscopy for analysis of chemical and microbiological properties of forest soil organic horizons in a heavy-metal-polluted area. Biol. Fert. Soils 44(1), 171-180. https://doi.org/10.1007/s00374-007-0192-z.
Christensen, T.H., Haung, P.M., 1999. Solid phase cadmium and the reactions of aqueous cadmium with soil surfaces. In: M.J. McLaughlin, B.R. Singh (Eds.), Cadmium in Soils and Plants. Springer Netherlands, Dordrecht, pp. 65-96. https://doi.org/10.1007/978-94-011-4473-5_4.
Cipullo, S., Nawar, S., Mouazen, A.M., Campo-Moreno, P., Coulon, F., 2019. Predicting bioavailability change of complex chemical mixtures in contaminated soils using visible and near-infrared spectroscopy and random forest regression. Sci. Rep. 9(1), 4492. https://doi.org/10.1038/s41598-019-41161-w.

Clark, R.D., 1997. OptiSim: An extended dissimilarity selection method for finding diverse representative subsets. J. Chem. Inform. Comput. Sci. 37(6), 1181-1188. https://doi.org/10.1021/ci970282v.
Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy - quantitative-analysis techniques for remote-sensing applications. J. Geophys. Res. 89(Nb7), 6329-6340. https://doi.org/10.1029/JB089iB07p06329.
Clemens, S., Aarts, M.G.M., Thomine, S., Verbruggen, N., 2013. Plant science: the key to preventing slow cadmium poisoning. Trends Plant Sci. 18(2), 92-99. https://doi.org/10.1016/j.tplants.2012.08.003.
Clingensmith, C.M., Grunwald, S., Wani, S.P., 2019. Evaluation of calibration subsetting and new chemometric methods on the spectral prediction of key soil properties in a data-limited environment. Eur. J. Soil Sci. 70(1), 107-126. https://doi.org/10.1111/ejss.12753.
Clyde, M., 1999. Bayesian model averaging and model search strategies. In: J. Bernardo, J. Berger, A. Dawid, A. Smith (Eds.), Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting. Oxford University Press, pp. 157-185.
Clyde, M., 2020. BAS: Bayesian variable selection and model averaging using bayesian adaptive sampling, R package version 1.5.5.
Cottenie, A., Verloo, M., 1984. Analytical diagnosis of soil pollution with heavy metals, Environmental Research and Protection. Environmental Research and Protection. Springer Berlin, Heidelberg, pp. 389-393.
Cramer, R.D., 1993. Partial least squares (PLS): its strengths and limitations. Perspec. Drug Discover. Design 1(2), 269-278. https://doi.org/10.1007/BF02174528.
Crippen, R.E., 1990. Calculating the vegetation index faster. Remote Sens. Environ. 34(1), 71-73. https://doi.org/10.1016/0034-4257(90)90085-Z.
Crush, J.R., Ouyang, L., Cousins, G.R., 2019. Variation in cadmium concentrations in shoots of chicory (Cichorium intybus L.). N. Z. J. Agric. Res. 62(4), 495-503. https://doi.org/10.1080/00288233.2018.1517806.
Cui, S., Zhou, K., Ding, R., Wang, J., Cheng, Y., Jiang, G., 2021. Monitoring the soil copper pollution degree based on the reflectance spectrum of an arid desert plant. Spectrochim. Acta A Mol. Biomol. Spectrosc. 263, 120186. https://doi.org/10.1016/j.saa.2021.120186.
Cunha, C.L., Torres, A.R., Luna, A.S., 2020. Multivariate regression models obtained from near-infrared spectroscopy data for prediction of the physical properties of biodiesel and its blends. Fuel 261, 116344. https://doi.org/10.1016/j.fuel.2019.116344.
CWG, 2011. Cadmium and New Zealand agriculture and horticulture: a strategy for long term risk management A report prepared by the Cadmium Working Group for the Chief Executives Environmental Forum. MAF Technical Paper No.: 2011/02 (Accessed 24 August 2018).
Dai, L.-P., Xiong, Z.-T., Huang, Y., Li, M.-J., 2006. Cadmium-induced changes in pigments, total phenolics, and phenylalanine ammonia-lyase activity in fronds ofAzolla imbricata. Environ. Toxic. 21(5), 505-512. https://doi.org/10.1002/tox.20212.
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(1), 11. https://doi.org/10.3390/soilsystems3010011.
Datt, B., 1998. Remote sensing of chlorophyll a, chlorophyll b, chlorophyll a+b, and total carotenoid content in eucalyptus leaves. Remote Sens. Environ. 66(2), 111-121. https://doi.org/10.1016/S0034-4257(98)00046-7.
Datt, B., 1999. Visible/near infrared reflectance and chlorophyll content in eucalyptus leaves. Int. J. Remote Sens. 20(14), 2741-2759. https://doi.org/10.1080/014311699211778.
Daughtry, C.S.T., 2001. Discriminating crop residues from soil by shortwave infrared reflectance. Agron. J. 93(1), 125-131. https://doi.org/10.2134/agronj2001.931125x.
Daughtry, C.S.T., Walthall, C.L., Kim, M.S., de Colstoun, E.B., McMurtrey, J.E., 2000. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. Remote Sens. Environ. 74(2), 229-239. https://doi.org/10.1016/S0034-4257(00)00113-9.
de Caritat, P., Cooper, M., 2016. A continental-scale geochemical atlas for resource exploration and environmental management: The national geochemical survey of Australia. Geochem. : Explor., Environ., Anal. 16, 3-13. https://doi.org/10.1144/geochem2014-322.
Demattê, J.A.M., Campos, R.C., Alves, M.C., Fiorio, P.R., Nanni, M.R., 2004. Visible–NIR reflectance: a new approach on soil evaluation. Geoderma 121(1-2), 95-112. https://doi.org/10.1016/j.geoderma.2003.09.012.
Demattê, J.A.M., Dotto, A.C., Paiva, A.F.S., Sato, M.V., Dalmolin, R.S.D., de Araújo, M.d.S.B., da Silva, E.B., Nanni, M.R., ten Caten, A., Noronha, N.C., Lacerda, M.P.C., de Araújo Filho, J.C., Rizzo, R., Bellinaso, H., Francelino, M.R., Schaefer, C.E.G.R., Vicente, L.E., dos Santos, U.J., de Sá Barretto Sampaio, E.V., Menezes, R.S.C., de Souza, J.J.L.L., Abrahão, W.A.P., Coelho, R.M.,

Grego, C.R., Lani, J.L., Fernandes, A.R., Gonçalves, D.A.M., Silva, S.H.G., de Menezes, M.D., Curi, N., Couto, E.G., dos Anjos, L.H.C., Ceddia, M.B., Pinheiro, É.F.M., Grunwald, S., Vasques, G.M., Marques Júnior, J., da Silva, A.J., Barreto, M.C.d.V., Nóbrega, G.N., da Silva, M.Z., de Souza, S.F., Valladares, G.S., Viana, J.H.M., da Silva Terra, F., Horák-Terra, I., Fiorio, P.R., da Silva, R.C., Frade Júnior, E.F., Lima, R.H.C., Alba, J.M.F., de Souza Junior, V.S., Brefin, M.D.L.M.S., Ruivo, M.D.L.P., Ferreira, T.O., Brait, M.A., Caetano, N.R., Bringhenti, I., de Sousa Mendes, W., Safanelli, J.L., Guimarães, C.C.B., Poppiel, R.R., e Souza, A.B., Quesada, C.A., do Couto, H.T.Z., 2019. The Brazilian soil spectral library (BSSL): A general view, application and challenges. Geoderma 354, 113793. https://doi.org/10.1016/j.geoderma.2019.05.043.

Dong, Y.-W., Yang, S.-Q., Xu, C.-Y., Li, Y.-Z., Bai, W., Fan, Z.-N., Wang, Y.-N., Li, Q.-Z., 2011. Determination of soil parameters in apple-growing regions by near- and mid-infrared spectroscopy. Pedosphere 21(5), 591-602. https://doi.org/10.1016/s1002-0160(11)60161-6.

Douglas, R.K., Nawar, S., Alamar, M.C., Coulon, F., Mouazen, A.M., 2019. The application of a handheld mid-infrared spectrometry for rapid measurement of oil contamination in agricultural sites. Sci. Total Environ. 665, 253-261. https://doi.org/10.1016/j.scitotenv.2019.02.065.

Drewry, J.J., Cavanagh, J.A.E., McNeill, S.J., Stevenson, B.A., Gordon, D.A., Taylor, M.D., 2021. Long-term monitoring of soil quality and trace elements to evaluate land use effects and temporal change in the Wellington region, New Zealand. Geoderma Reg. 25, e00383. https://doi.org/10.1016/j.geodrs.2021.e00383.

Du Laing, G., 2010. Analysis and fractionation of trace elements in soils. In: P.S. Hooda (Ed.), Trace elements in soils. Blackwell publishing, pp. 53-80. https://doi.org/10.1002/9781444319477.ch4.

El-Mufleh Al Husseini, A., Béchet, B., Gaudin, A., Ruban, V., 2013. Trace metal fractionation as a mean to improve on the management of contaminated sediments from runoff water in infiltration basins. Environ. Technol. 34(10), 1255-1266. https://doi.org/10.1080/09593330.2012.745619.

Elam, W.T., Ravel, B.D., Sieber, J.R., 2002. A new atomic database for X-ray spectroscopic calculations. Radiat. Phys. Chem. 63(2), 121-128. https://doi.org/10.1016/S0969-806x(01)00227-4.

Fan, P., Wu, L., Wang, Q., Wang, Y., Luo, H., Song, J., Yang, M., Yao, H., Chen, S., 2023. Physiological and molecular mechanisms of medicinal plants in response to cadmium stress: Current status and future perspective. J. Hazard. Mater. 450, 131008. https://doi.org/10.1016/j.jhazmat.2023.131008.

FAO/WHO, 2011. Evaluation of certain food additives and contaminants: Seventy-third report of the Joint FAO/WHO expert committee on food additives.

Farrés, M., Platikanov, S., Tsakovski, S., Tauler, R., 2015. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. J. Chemometric. 29(10), 528-536. https://doi.org/10.1002/cem.2736.

Feng, X., Chen, H., Chen, Y., Zhang, C., Liu, X., Weng, H., Xiao, S., Nie, P., He, Y., 2019. Rapid detection of cadmium and its distribution in Miscanthus sacchariflorus based on visible and near-infrared hyperspectral imaging. Sci. Total Environ. 659, 1021-1031. https://doi.org/10.1016/j.scitotenv.2018.12.458.

Fertiliser Association, 2016. Tiered Fertiliser Management System: for the management of soil cadmium accumulation from phosphate fertiliser applications, Developed by the Fertiliser Association of New Zealand as part of the work programme undertaken for the Cadmium Working Group, which was established by the Regional Councils' Chief Executives' Environment Forum. version 2.

Fertiliser Association, 2019. Tiered fertiliser management system: For the management of soil cadmium accumulation from phosphate fertiliser applications, Developed by the Fertiliser Association of New Zealand as part of the work programme undertaken for the Cadmium Working Group, which was established by the Regional Councils' Chief Executives' Environment Forum. Version 3.

Förstner, U., 1993. Metal Speciation - General Concepts and Applications. Int. J. Environ. Anal. Chem. 51(1-4), 5-23. https://doi.org/10.1080/03067319308027608. .

FSANZ, 2016. Australia New Zealand Food Standards Code - Schedule 19 - Maximum levels of contaminant and natural toxicants.

Fu, P., Zhang, W., Yang, K., Meng, F., 2020. A novel spectral analysis method for distinguishing heavy metal stress of maize due to copper and lead: RDA and EMD-PSD. Ecotox. Environ. Saf. 206, 111211. https://doi.org/10.1016/j.ecoenv.2020.111211.

Gamon, J.A., Peñuelas, J., Field, C.B., 1992. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. Remote Sens. Environ. 41(1), 35-44. https://doi.org/10.1016/0034-4257(92)90059-S.

Gamon, J.A., Surfus, J.S., 1999. Assessing leaf pigment content and activity with a reflectometer. New Phytol. 143(1), 105-117. https://doi.org/10.1046/j.1469-8137.1999.00424.x.

Gazley, M.F., Martin, A.P., Turnbull, R.E., Frontin-Rollet, G., Strong, D.T., 2020. Regional patterns in standardised and transformed pathfinder elements in soil related to orogenic-style mineralisation

in southern New Zealand. J. Geochem. Explor. 217, 106593. https://doi.org/10.1016/j.gexplo.2020.106593.
Ghamisi, P., Yokoya, N., Li, J., Liao, W., Liu, S., Plaza, J., Rasti, B., Plaza, A., 2017. Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. IEEE Geosci. Remote Sens. Mag. 5(4), 37-78. https://doi.org/10.1109/mgrs.2017.2762087.
Gholizadeh, A., Boruvka, L., Vasat, R., Saberioon, M., Klement, A., Kratina, J., Tejnecky, V., Drabek, O., 2015. Estimation of potentially toxic elements contamination in anthropogenic soils on a brown coal mining dumpsite by reflectance spectroscopy: A case study. PLOS ONE 10(2), e0117457. https://doi.org/10.1371/journal.pone.0117457.
Gholizadeh, A., Coblinski, J.A., Saberioon, M., Ben-Dor, E., Drabek, O., Dematte, J.A.M., Boruvka, L., Nemecek, K., Chabrillat, S., Dajcl, J., 2021. vis-NIR and XRF data fusion and feature selection to estimate potentially toxic elements in soil. Sensors (Basel, Switzerland) 21(7). https://doi.org/10.3390/s21072386.
Gholizadeh, A., Kopackova, V., 2019. Detecting vegetation stress as a soil contamination proxy: A review of optical proximal and remote sensing techniques. Int. J. Environ. Sci. Tech. 16(5), 2511-2524. https://doi.org/10.1007/s13762-019-02310-w.
Gitelson, A., Merzlyak, M.N., 1994. Spectral reflectance changes associated with autumn senescence of Aesculus hippocastanum L. and Acer platanoides L. leaves spectral features and relation to chlorophyll estimation. J. Plant Physiol. 143(3), 286-292. https://doi.org/10.1016/S0176-1617(11)81633-0.
Gitelson, A.A., Keydan, G.P., Merzlyak, M.N., 2006. Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. Geophys. Res. Letters 33(11). https://doi.org/10.1029/2006gl026457.
Gitelson, A.A., Merzlyak, M.N., 1997. Remote estimation of chlorophyll content in higher plant leaves. Int. J. Remote Sens. 18(12), 2691-2697. https://doi.org/10.1080/014311697217558.
Gitelson, A.A., Merzlyak, M.N., Chivkunova, O.B., 2001. Optical properties and nondestructive estimation of anthocyanin content in plant leaves. Photochem. Photobiol. 74(1), 38-45. https://doi.org/10.1562/0031-8655(2001)0740038opaneo2.0.co2.
Gleyzes, C., Tellier, S., Astruc, M., 2002. Fractionation studies of trace elements in contaminated soils and sediments: A review of sequential extraction procedures. Trac-Trend Anal. Chem. 21(6-7), 451-467. https://doi.org/10.1016/s0165-9936(02)00603-9.
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(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.
Gomez, C., Chevallier, T., Moulin, P., Bouferra, I., Hmaidi, K., Arrouays, D., Jolivet, C., Barthes, B.G., 2020. Prediction of soil organic and inorganic carbon concentrations in Tunisian samples by mid-infrared reflectance spectroscopy using a French national library. Geoderma 375, 114469. https://doi.org/10.1016/j.geoderma.2020.114469.
Götze, C., Jung, A., Merbach, I., Wennrich, R., Gläßer, C., 2010. Spectrometric analyses in comparison to the physiological condition of heavy metal stressed floodplain vegetation in a standardised experiment. Cent. Eur. J. Geosci. 2(2), 132-137. https://doi.org/10.2478/v10085-010-0002-y.
Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. J. Forecast. 3(2), 197-204. https://doi.org/10.1002/for.3980030207.
Gratão, P.L., Gomes-Junior, R.A., Delite, F.S., Lea, P.J., Azevedo, R.A., 2006. Antioxidant stress responses of plants to cadmium. In: N.A. Khan, Samiullah (Eds.), Cadmium toxicity and tolerance in plants. Alpha Science Int. Ltd., Oxford, pp. 1-34.
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., 1-51. https://doi.org/10.1080/00288233.2022.2069130.
Gray, C.W., McLaren, R.G., Roberts, A.H.C., Condron, L.M., 1998. Sorption and desorption of cadmium from some New Zealand soils: Effect of pH and contact time. Soil Res. 36(2), 199-216. https://doi.org/10.1071/S97085.
Gray, C.W., McLaren, R.G., Roberts, A.H.C., Condron, L.M., 1999a. Cadmium phytoavailability in some New Zealand soils. Aust. J. Soil Res. 37(3), 461-478. https://doi.org/10.1071/s98070.
Gray, C.W., McLaren, R.G., Roberts, A.H.C., Condron, L.M., 1999b. The effect of long-term phosphatic fertiliser applications on the amounts and forms of cadmium in soils under pasture in New Zealand. Nutr. Cycl. Agroecosys. 54(3), 267-277. https://doi.org/10.1023/A:1009883010490.

Gray, C.W., McLaren, R.G., Roberts, A.H.C., Condron, L.M., 1999c. Solubility, sorption and desorption of native and added cadmium in relation to properties of soils in New Zealand. Eur. J. Soil Sci. 50(1), 127-137. https://doi.org/10.1046/j.1365-2389.1999.00221.x.
Gray, C.W., McLaren, R.G., Roberts, A.H.C., Condron, L.M., 2000. Fractionation of soil cadmium from some New Zealand soils. Commun. Soil Sci. Plant Anal. 31(9-10), 1261-1273. https://doi.org/10.1080/00103620009370511.
Gray, C.W., Yi, Z., Lehto, N.J., Robinson, B.H., Munir, K., Cavanagh, J.-A.E., 2019a. Effect of cultivar type and soil properties on cadmium concentrations in potatoes. N. Z. J. Crop Hort. Sci. 47(3), 182-197. https://doi.org/10.1080/01140671.2019.1599028.
Gray, C.W., Yi, Z., Munir, K., Lehto, N.J., Robinson, B.H., Cavanagh, J.E., 2019b. Cadmium concentrations in New Zealand wheat: Effect of cultivar type, soil properties, and crop management. J. Environ. Qual. 48(3), 701-708. https://doi.org/10.2134/jeq2018.12.0430.
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.
Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., Goodacre, R., 2015. A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. Anal. Chim. Acta. 879, 10-23. https://doi.org/10.1016/j.aca.2015.02.012.
Grove, C., Hook, S.J., Paylor III, E., 1992. Laboratory reflectance spectra of 160 minerals, 0.4 to 2.5 micrometers, Pasadena, CA: Jet Propulsion Laboratory.
Guérin, T., Ghinet, A., Waterlot, C., 2022. The phytoextraction power of Cichorium intybus L. on metal-contaminated soil: Focus on time- and cultivar-depending accumulation and distribution of cadmium, lead and zinc. Chemosphere 287, 132122. https://doi.org/10.1016/j.chemosphere.2021.132122.
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(1-2), 66-77. https://doi.org/10.1016/j.geoderma.2009.12.021.
Guo, Q., Wu, W., Massart, D.L., 1999. The robust normal variate transform for pattern recognition with near-infrared data. Anal. Chim. Acta. 382(1-2), 87-103. https://doi.org/10.1016/S0003-2670(98)00737-5.
Haboudane, D., Miller, J.R., Pattey, E., Zarco-Tejada, P.J., Strachan, I.B., 2004. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. Remote Sens. Environ. 90(3), 337-352. https://doi.org/10.1016/j.rse.2003.12.013.
Haboudane, D., Miller, J.R., Tremblay, N., Zarco-Tejada, P.J., Dextraze, L., 2002. Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. Remote Sens. Environ. 81(2), 416-426. https://doi.org/10.1016/S0034-4257(02)00018-4.
Han, H., Jiang, X., 2014. Overcome support vector machine diagnosis overfitting. Cancer Inform. 13(Suppl 1), 145-158. https://doi.org/10.4137/cin.s13875.
Hao, Z., Du, J., Nie, B., Yu, F., Yu, R., Xiong, W., 2016. Random forest regression based on partial least squares connect partial least squares and random forest, 2016 International Conference on Artificial Intelligence: Technologies and Applications. Atlantis Press.
Harrell, F., DuPont, C., 2021. Hmisc: Harrell Miscellaneous. R package version 4.5-0.
Harvey, P.J., Rouillon, M., Dong, C., Ettler, V., Handley, H.K., Taylor, M.P., Tyson, E., Tennant, P., Telfer, V., Trinh, R., 2017. Geochemical sources, forms and phases of soil contamination in an industrial city. Sci. Total Environ. 584-585, 505-514. https://doi.org/10.1016/j.scitotenv.2017.01.053.
Hewitt, A., Balks, M.R., Lowe, D.J., 2021. The Soils of Aotearoa New Zealand. World Soils Book Series. Springer. https://doi.org/10.1007/978-3-030-64763-6.
Hewitt, A.E., 2010. New Zealand soil classification. 3rd ed. Manaaki-Whenua - Landcare Research New Zealand Ltd.
Hong, Y.S., Chen, Y.Y., Zhang, Y., Liu, Y.F., Liu, Y.L., Yu, L., Liu, Y., Cheng, H., 2018. Transferability of Vis-NIR models for soil organic carbon estimation between two study areas by using spiking. Soil Sci. Soc. Am. J. 82(5), 1231-1242. https://doi.org/10.2136/sssaj2018.03.0099.
Hong, Y.S., Liu, Y., Chen, Y.Y., Liu, Y.F., Yu, L., Liu, Y., Cheng, H., 2019. Application of fractional-order derivative in the quantitative estimation of soil organic matter content through visible and near-infrared spectroscopy. Geoderma 337, 758-769. https://doi.org/10.1016/j.geoderma.2018.10.025.
Hooda, P.S. (Ed.), 2010. Trace elements in soils. A John Wiley & Sons Ltd Publication, United Kingdom, 618 pp. https://doi.org/10.1002/9781444319477.

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25(3), 295-309. https://doi.org/10.1016/0034-4257(88)90106-X.
Hunt, G.R., Salisbury, J., Lenhoff, C., 1974. Visible and near infrared spectra of minerals and rocks: Basic and ultrabasic igneous rocks. Mod. Geol. 5, 15-22.
Hunt, G.R., Slisbury, J., 1970. Visible and near-infrared spectra of minerals and rocks: I. silicate minerals. Mod. Geol. 1, 283-300.
Hutengs, C., Eisenhauer, N., Schädler, M., Lochner, A., Seidel, M., Vohland, M., 2021. VNIR and MIR spectroscopy of PLFA-derived soil microbial properties and associated soil physicochemical characteristics in an experimental plant diversity gradient. Soil Biol. Biochem. 160, 108319. https://doi.org/10.1016/j.soilbio.2021.108319.
IUSS Working Group WRB, 2015. World Reference Base for Soil Resources 2014, update 2015 International soil classification system for naming and creating legends for soil maps. World Soil Resources Reports No. 106. FAO, Rome.
Jaillais, B., Pinto, R., Barros, A.S., Rutledge, D.N., 2005. Outer-product analysis (OPA) using PCA to study the influence of temperature on NIR spectra of water. Vibrat. Spectrosc. 39(1), 50-58. https://doi.org/10.1016/j.vibspec.2004.10.008.
Janik, L.J., Merry, R.H., Skjemstad, J.O., 1998. Can mid infrared diffuse reflectance analysis replace soil extractions? Aust. J. Exp. Agric. 38(7), 681-696. https://doi.org/10.1071/ea97144.
Järup, L., Åkesson, A., 2009. Current status of cadmium as an environmental health problem. Toxicol. Appl. Pharmacol. 238(3), 201-208. https://doi.org/10.1016/j.taap.2009.04.020.
Javadi, S.H., Munnaf, M.A., Mouazen, A.M., 2021. Fusion of Vis-NIR and XRF spectra for estimation of key soil attributes. Geoderma 385, 114851. https://doi.org/10.1016/j.geoderma.2020.114851.
Jeyakumar, P., Loganathan, P., Sivakumaran, S., Anderson, C.W.N., McLaren, R.G., 2010. Bioavailability of copper and zinc to poplar and microorganisms in a biosolids-amended soil. Aust. J. Soil. Res. 48(5), 459-469. https://doi.org/10.1071/Sr09169.
Jiang, B., Su, D.C., Wang, X.Q., Liu, J.F., Ma, Y.B., 2018. Field evidence of decreased extractability of copper and nickel added to soils in 6-year field experiments. Front. Env. Sci. Eng. 12(2), 7. https://doi.org/10.1007/s11783-017-0990-y.
Jiang, Q., Li, Q., Wang, X., Wu, Y., Yang, X., Liu, F., 2017. Estimation of soil organic carbon and total nitrogen in different soil layers using VNIR spectroscopy: Effects of spiking on model applicability. Geoderma 293, 54-63. https://doi.org/10.1016/j.geoderma.2017.01.030.
Jones, R.P., Hassan, S.M., Rodgers, J.H., Jr., 2008. Influence of contact duration on sediment-associated copper fractionation and bioavailability. 71(1), 104-116. 10.1016/j.ecoenv.2007.09.004.
Jordan, C.F., 1969. Derivation of leaf-area index from quality of light on the forest floor. Ecol. 50(4), 663-666. https://doi.org/10.2307/1936256.
Kabata-Pendias, A., 2010. Trace elements in soils and plants. Fourth ed. CRC Press. https://doi.org/10.1201/b10158.
Kalnicky, D.J., Singhvi, R., 2001. Field portable XRF analysis of environmental samples. J. Hazard. Mater. 83(1-2), 93-122. https://doi.org/10.1016/s0304-3894(00)00330-7.
Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. kernlab – An S4 package for kernel methods in R. J. Stat. Softw. 11(9), 1–20. http://www.jstatsoft.org/v11/i09/.
Kashem, M.A., Singh, B.R., Huq, S.I., Kawai, S., 2011. Fractionation and mobility of cadmium, lead and zinc in some contaminated and non-contaminated soils of Japan. J. Soil Sci. Environ. Manag. 2(9), 241-249.
Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J.N., Pachepsky, Y., van der Putten, W.H., Bardgett, R.D., Moolenaar, S., Mol, G., Jansen, B., Fresco, L.O., 2016. The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. SOIL 2(2), 111-128. https://doi.org/10.5194/soil-2-111-2016.
Kemper, T., Sommer, S., 2002. Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. Environ. Sci. Technol. 36(12), 2742-2747. https://doi.org/10.1021/es015747j.
Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11(1), 137-148. https://doi.org/10.2307/1266770.
Khosravi, V., Ardejani, F.D., Yousefi, S., Aryafar, A., 2018. Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods. Geoderma 318, 29-41. https://doi.org/10.1016/j.geoderma.2017.12.025.
Kim, B., 2006. The long-term behavior of trace metals applied to soils at toxic levels . Doctoral Thesis, Cornell University, US, 109 pp.
Kim, N., 2005. Cadmium accumulation in Waikato soils. 2005/51, Environment Waikato Regional Council, Hamilton, New Zealand.

Kim, N., Fergusson, J., 1991. Effectiveness of a commonly used sequential extraction technique in determining the speciation of cadmium in soils. Sci. Total Environ. 105, 191-209. https://doi.org/10.1016/0048-9697(91)90341-b.
Kooistra, L., Salas, E.A.L., Clevers, J.G.P.W., Wehrens, R., Leuven, R.S.E.W., Nienhuis, P.H., Buydens, L.M.C., 2004. Exploring field vegetation reflectance as an indicator of soil contamination in river floodplains. Environ. Pollut. 127(2), 281-290. https://doi.org/10.1016/S0269-7491(03)00266-5.
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(4), 337-344. https://doi.org/10.1016/s0303-2434(01)85041-8.
Kovács, B., Prokisch, J., Györi, Z., Kovács, A.B., Palencsár, A.J., 2000. Studies on soil sample preparation for inductively coupled plasma atomic emission spectrometry analysis. Comm. Soil Sci. Plant Anal. 31(11-14), 1949-1963. https://doi.org/10.1080/00103620009370553.
Krishnamurti, G.S., 2008. Chemical methods for assessing contaminant bioavailability in soils. In: R. Naidu (Ed.), Chemical bioavailability in terrestrial environments. Elsevier, pp. 495-520. https://doi.org/10.1016/S0166-2481(07)32020-5.
Krishnamurti, G.S.R., Huang, P.M., Vanrees, K.C.J., Kozak, L.M., Rostad, H.P.W., 1995. Speciation of particulate-bound cadmium of soils and its bioavailability. Analyst 120(3), 659-665. https://doi.org/10.1039/an9952000659.
Krishnan, P., Alexander, J.D., Butler, B.J., Hummel, J.W., 1980. Reflectance technique for predicting soil organic matter. Soil Sci. Soc. Am. J. 44(6), 1282-1285. https://doi.org/10.2136/sssaj1980.03615995004400060030x.
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 Till. Res. 128, 125-136. https://doi.org/10.1016/j.still.2012.11.006.
Kubier, A., Wilkin, R.T., Pichler, T., 2019. Cadmium in soils and groundwater: A review. Appl. Geochem. 108, 1-16. https://doi.org/10.1016/j.apgeochem.2019.104388.
Kuhn, M., Johnson, K., 2013. Applied predictive modeling, 26. 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.
Lai, H.Y., Chen, Z.S., 2006. Phytoremidiation techniques of Cd-contaminated soils: toxicity, enhanced uptake techniques, and mechanisms. In: N.A. Khan, Samiullah (Eds.), Cadmium toxicity and tolerance in plants. Alpha Science International Ltd., Oxford, U.K., pp. 35-61.
Lamine, S., Petropoulos, G.P., Brewer, P.A., Bachari, N.E., Srivastava, P.K., Manevski, K., Kalaitzidis, C., Macklin, M.G., 2019. Heavy metal soil contamination detection using combined geochemistry and field spectroradiometry in the United Kingdom. Sensors (Basel) 19(4), 762. https://doi.org/10.3390/s19040762.
Largo-Gosens, A., Hernández-Altamirano, M., García-Calvo, L., Alonso-Simón, A., Álvarez, J., Acebes, J.L., 2014. Fourier transform mid infrared spectroscopy applications for monitoring the structural plasticity of plant cell walls. Frontiers Plant Sci. 5(303). https://doi.org/10.3389/fpls.2014.00303.
Lee, J., Rounce, J., Mackay, A., Grace, N., 1996. Accumulation of cadmium with time in Romney sheep grazing ryegrass-white clover pasture: Effect of cadmium from pasture and soil intake. Aus. J. Agric. Res. 47(6), 877. https://doi.org/10.1071/ar9960877.
Lehmann, J., Bossio, D.A., Kögel-Knabner, I., Rillig, M.C., 2020. The concept and future prospects of soil health. Nat. Rev. Earth Environ. 1(10), 544-553. https://doi.org/10.1038/s43017-020-0080-8.
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, F., Lu, A., Wang, J., 2017. Modeling of chromium, copper, zinc, arsenic and lead using portable X-ray fluorescence spectrometer based on discrete wavelet transform. Int. J. Environ. Res. Pub. Health 14(10), 1163. https://doi.org/10.3390/ijerph14101163.
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.
Li, G., Kemp, P.D., 2005. Forage chicory (*Cichorium intybus* L.): A review of its agronomy and animal production, Advances in Agronomy. Academic Press, pp. 187-222. https://doi.org/10.1016/S0065-2113(05)88005-8.
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(16), 4357. https://doi.org/10.3390/s20164357.

Li, M., Xi, X.H., Xiao, G.Y., Cheng, H.X., Yang, Z.F., Zhou, G.H., Ye, J.Y., Li, Z.H., 2014. National multi-purpose regional geochemical survey in China. J. Geochem. Explor. 139, 21-30. https://doi.org/10.1016/j.gexplo.2013.06.002.

Lichtenthaler, H.K., Lang, M., Sowinska, M., Heisel, F., Miehé, J.A., 1996. Detection of vegetation stress via a new high resolution fluorescence imaging system. J. Plant Physiol. 148(5), 599-612. https://doi.org/10.1016/S0176-1617(96)80081-2.

Lin, L.I., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45(1), 255-268. https://doi.org/10.2307/2535051.

Liu, K., Zhao, D., Fang, J.-y., Zhang, X., Zhang, Q.-y., Li, X.-k., 2017. Estimation of heavy-metal contamination in soil using remote sensing spectroscopy and a statistical approach. J. Indian Soc. Remote Sens. 45(5), 805-813. https://doi.org/10.1007/s12524-016-0648-4.

Liu, Y., Wang, C., Xiao, C., Shang, K., Zhang, Y., Pan, X., 2021. Prediction of multiple soil fertility parameters using VisNIR spectroscopy and PXRF spectrometry. Soil Sci. Soc. Am. J. 85(3), 591-605. https://doi.org/0.1002/saj2.20223.

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

Loganathan, P., Louie, K., Lee, J., Hedley, M.J., Roberts, A.H.C., Longhurst, R.D., 1999. A model to predict kidney and liver cadmium concentrations in grazing animals. N. Z. J. Agric. Res. 42(4), 423-432. https://doi.or/g10.1080/00288233.1999.9513391.

Loganathan, P., Vigneswaran, S., Kandasamy, J., Naidu, R., 2012. Cadmium sorption and desorption in soils: A review. Crit. Rev. Environ. Sci. Technol. 42(5), 489-533. https://doi.org/10.1080/10643389.2010.520234.

Longhurst, R.D., 2006. Envirolink 73 - HBRC 9 - soil cadmium: Report prepared for Hawkes Bay Regional Council.

Longhurst, R.D., Roberts, A.H.C., Waller, J.E., 2004. Concentrations of arsenic, cadmium, copper, lead, and zinc in New Zealand pastoral topsoils and herbage. N. Z. J. Agric. Res. 47(1), 23-32. https://doi.org/10.1080/00288233.2004.9513567.

MacLeod, C.J., Moller, H., 2006. Intensification and diversification of New Zealand agriculture since 1960: An evaluation of current indicators of land use change. Agric. Ecosys. Environ. 115(1), 201-218. https://doi.org/10.1016/j.agee.2006.01.003.

Mahlein, A.K., Rumpf, T., Welke, P., Dehne, H.W., Plümer, L., Steiner, U., Oerke, E.C., 2013. Development of spectral indices for detecting and identifying plant diseases. Remote Sens. Environ. 128, 21-30. https://doi.org/10.1016/j.rse.2012.09.019.

Malone, B., Hedley, C., Roudier, P., Minasny, B., Jones, E., McBratney, A., 2018. Auditing on-farm soil carbon stocks using downscaled national mapping products: Examples from Australia and New Zealand. Geoderma Reg. 13, 1-14. https://doi.org/10.1016/j.geodrs.2018.02.002.

Mann, S.S., Rate, A.W., 1998. Determination of cadmium in soil extracts containing high levels of iron and aluminum by graphite furnace atomic absorption spectrophotometry. Commun. Soil Sci. Plant Anal. 29(17-18), 2725-2737. https://doi.org/10.1080/00103629809370147.

Marković, J., Jović, M., Smičiklas, I., Šljivić-Ivanović, M., Onjia, A., Trivunac, K., Popović, A., 2019. Cadmium retention and distribution in contaminated soil: effects and interactions of soil properties, contamination level, aging time and in situ immobilization agents. Ecotoxicol. Environ. Saf. 174, 305-314. https://doi.org/10.1016/j.ecoenv.2019.03.001.

Marschner, H., 2012. Marschner's mineral nutrition of higher plants. Third ed. Academic Press. https://doi.org/10.1016/c2009-0-63043-9.

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.

Martin, A.P., Turnbull, R.E., Rissmann, C.W., Rieger, P., 2017. Heavy metal and metalloid concentrations in soils under pasture of southern New Zealand. Geoderma Reg. 11, 18-27. https://doi.org/10.1016/j.geodrs.2017.08.005.

Massart, D.L., Vandeginste, B.G., Buydens, L., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J., 1997. Handbook of chemometrics and qualimetrics: Part A. Data Handling in Science and Technology, 20A.

Mattina, M.I., Lannucci-Berger, W., Musante, C., White, J.C., 2003. Concurrent plant uptake of heavy metals and persistent organic pollutants from soil. Environ. Pollut. 124(3), 375-378. https://doi.org/10.1016/S0269-7491(03)00060-5.

McBride, M.B., Li, X., 2022. Toxic metal persistence and bioavailability in agricultural soil 40 years after sewage sludge incorporation. J. Soils Sediments. https://doi.org/10.1007/s11368-022-03314-9.
McCarty, G.W., Reeves, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M., 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. Soil Sci. Soc. Am. J. 66(2), 640-646. https://doi.org/10.2136/sssaj2002.0640.
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.
McDowell, R.W., Taylor, M.D., Stevenson, B.A., 2013. Natural background and anthropogenic contributions of cadmium to New Zealand soils. Agric. Ecosys. Environ. 165, 80-87. https://doi.org/10.1016/j.agee.2012.12.011.
McGladdery, C., Weindorf, D.C., Chakraborty, S., Li, B., Paulette, L., Podar, D., Pearson, D., Kusi, N.Y.O., Duda, B., 2018. Elemental assessment of vegetation via portable X-ray fluorescence (PXRF) spectrometry. 210, 210-225. 10.1016/j.jenvman.2018.01.003.
McKenzie, H.A., Wallace, H.S., 1954. The Kjeldahl determination of nitrogen: A critical study of digestion conditions - temperature, catalyst, and oxidizing agent. Aust. J. Chem. 7(1), 55-70. https://doi.org/10.1071/ch9540055.
McKone, T.E., Maddalena, R.L., 2007. Plant uptake of organic pollutants from soil: Bioconcentration estimates based on models and experiments. Environ. Toxicol. Chem. 26(12), 2494. https://doi.org/10.1897/06-269.1.
McLaren, T.I., Guppy, C.N., Tighe, M.K., 2012. A rapid and nondestructive plant nutrient analysis using portable X-ray fluorescence. Soil Sci. Soc. Am. J. 76(4), 1446-1453. https://doi.org/10.2136/sssaj2011.0355.
McMurtrey, J.E., Chappelle, E.W., Kim, M.S., Meisinger, J.J., Corp, L.A., 1994. Distinguishing nitrogen fertilization levels in field corn (Zea mays L.) with actively induced fluorescence and passive reflectance measurements. Remote Sens. Environ. 47(1), 36-44. https://doi.org/10.1016/0034-4257(94)90125-2.
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.
Merzlyak, M.N., Gitelson, A.A., Chivkunova, O.B., Rakitin, V.Y., 1999. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. Physiol. Plantarum 106(1), 135-141. https://doi.org/10.1034/j.1399-3054.1999.106119.x.
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.
MIEF, 2007. Government decree on the assessment of soil contamination and remediation needs (214/2007), Ministry of the Environment Helsinki (FI).
Minasny, B., McBratney, A., Stockmann, U., Hong, S., 2013. Cubist, a regression rule approach for use in calibration of NIR spectra. In: V.B. Maurel, P. Williams, G. Downey, R. Kabore (Eds.), Picking up Good Vibrations, France, pp. 630-633.
Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32(9), 1378-1388. https://doi.org/10.1016/j.cageo.2005.12.009.
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.
Ministry for the Environment, Ministry for Primary Industries, 2022. National policy statement for highly productive land 2022. https://environment.govt.nz/publications/national-policy-statement-for-highly-productive-land/, New Zealand, pp. 16.
Morgan, R. (Ed.), 2010. Soil, heavy metals, and human health. Soils and human health. CRC Press, 59-82 pp. https://doi.org/10.1201/b13683.
Moros, J., Fdez-Ortiz De Vallejuelo, S., Gredilla, A., De Diego, A., Madariaga, J.M., Garrigues, S., De La Guardia, M., 2009. Use of reflectance infrared spectroscopy for monitoring the metal content of the estuarine sediments of the Nerbioi-Ibaizabal River (Metropolitan Bilbao, Bay of Biscay, Basque Country). Environ. Sci. Technol. 43(24), 9314-9320. https://doi.org/10.1021/es9005898.
Morra, M.J., Hall, M.H., Freeborn, L.L., 1991. Carbon and nitrogen analysis of soil fractions using near-infrared reflectance spectroscopy. Soil Sci. Soc. Am. J. 55(1), 288-291. https://doi.org/10.2136/sssaj1991.03615995005500010051x.

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.

Naes, T., Isaksson, T., Kowalski, B., 2002. Locally weighted regression and scatter correction for near-infrared reflectance data. Anal. Chem. 62(7), 664-673. https://doi.org/10.1021/ac00206a003.

Naidu, R., Bolan, N.S., Kookana, R.S., Tiller, K.G., 1994. Ionic-strength and pH effects on the sorption of cadmium and the surface charge of soils. Eur. J. Soil Sci. 45(4), 419-429. https://doi.org/10.1111/j.1365-2389.1994.tb00527.x.

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(7), 525-557. https://doi.org/10.1080/05704928.2019.1608110.

Nawar, S., Mouazen, A., 2017a. Comparison between Random Forests, Artificial Neural Networks and Gradient Boosted Machines Methods of On-Line Vis-NIR Spectroscopy Measurements of Soil Total Nitrogen and Total Carbon. 17(10), 2428. https://doi.org/10.3390/s17102428.

Nawar, S., Mouazen, A.M., 2017b. 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., Mouazen, A.M., 2018. Optimal sample selection for measurement of soil organic carbon using on-line vis-NIR spectroscopy. Comput. Electron. Agric. 151, 469-477. https://doi.org/10.1016/j.compag.2018.06.042.

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(5), 531-541. https://doi.org/10.4141/CJSS08076.

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

Ng, W., Minasny, B., Jones, E., McBratney, A., 2022b. To spike or to localize? Strategies to improve the prediction of local soil properties using regional spectral library. Geoderma 406, 115501. https://doi.org/10.1016/j.geoderma.2021.115501.

Ng, W., Minasny, B., Malone, B., Filippi, P., 2018. In search of an optimum sampling algorithm for prediction of soil properties from infrared spectra. PeerJ 6, e5722. https://doi.org/10.7717/peerj.5722.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. Geoderma 352, 251-267. https://doi.org/10.1016/j.geoderma.2019.06.016.

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. Tech. 12(6), 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. In: D.L. Sparks (Ed.). Adv. Agron. Academic Press, pp. 139-159. https://doi.org/10.1016/bs.agron.2015.02.002.

Nogawa, K., Kobayashi, E., Okubo, Y., Suwazono, Y., 2004. Environmental cadmium exposure, adverse effects and preventive measures in Japan. Biomet. 17(5), 581-587. https://doi.org/10.1023/B:BIOM.0000045742.81440.9c.

Nriagu, J.O., Pacyna, J.M., 1988. Quantitative assessment of worldwide contamination of air, water and soils by trace metals. Nature 333(6169), 134-139. https://doi.org/10.1038/333134a0.

Nwafor, I.C., Shale, K., Achilonu, M.C., 2017. Chemical composition and nutritive benefits of chicory *Cichorium intybus* as an ideal complementary and/or alternative livestock feed supplement. Scientific World J. 2017, 1-11. https://doi.org/10.1155/2017/7343928.

O'Rourke, S.M., Minasny, B., Holden, N.M., McBratney, A.B., 2016a. Synergistic use of Vis-NIR, MIR, and XRF spectroscopy for the determination of soil geochemistry. Soil Sci. Soc. Am. J. 80(4), 888-899. https://doi.org/10.2136/sssaj2015.10.0361.

O'Rourke, S.M., Stockmann, U., Holden, N.M., McBratney, A.B., Minasny, B., 2016b. An assessment of model averaging to improve predictive power of portable vis-NIR and XRF for the determination of agronomic soil properties. Geoderma 279, 31-44. https://doi.org/10.1016/j.geoderma.2016.05.005.

Oliver, I., Merrington, G., McLaughlin, M., 2006. Copper partitioning among mineral and organic fractions in biosolids. Environ. Chem. 3(1), 48-52. https://doi.org/10.1071/EN05066.

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.

Palacio, S., Aitkenhead, M., Escudero, A., Montserrat-Martí, G., Maestro, M., Robertson, A.H.J., 2014. Gypsophile chemistry unveiled: fourier transform infrared (FTIR) spectroscopy provides new insight into plant adaptations to gypsum soils. PLOS ONE 9(9), e107285. https://doi.org/10.1371/journal.pone.0107285.

Paltseva, A.A., Deeb, M., Di Iorio, E., Circelli, L., Cheng, Z., Colombo, C., 2022. Prediction of bioaccessible lead in urban and suburban soils with Vis-NIR diffuse reflectance spectroscopy. Sci. Total Environ. 809, 151107. https://doi.org/10.1016/j.scitotenv.2021.151107.

Patle, A., Chouhan, D.S., 2013. SVM kernel functions for classification, 2013 International Conference on Advances in Technology and Engineering (ICATE). IEEE, pp. 1-9.

PenUelas, J., Filella, I., Lloret, P., MunOz, F., Vilajeliu, M., 1995. Reflectance assessment of mite effects on apple trees. Int. J. Remote Sens. 16(14), 2727-2733. https://doi.org/10.1080/01431169508954588.

Peñuelas, J., Gamon, J.A., Fredeen, A.L., Merino, J., Field, C.B., 1994. Reflectance indices associated with physiological changes in nitrogen- and water-limited sunflower leaves. Remote Sens. Environ. 48(2), 135-146. https://doi.org/10.1016/0034-4257(94)90136-8.

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(3), 501-519. https://doi.org/10.1002/saj2.20225.

Poggio, M., Roudier, P., Blaschek, M., Hedley, C., 2018. Integration of NIR on a multi-sensor platform to improve soil resource assessments. NIR news 29(5), 15-18. https://doi.org/10.1177/0960336018782035.

Powell, A., Kemp, P., Jaya, I.D., Osborne, M., 2007. Establishment, growth and development of plantain and chicory under grazing, Proceedings of the New Zealand Grassland Association, New Zealand, pp. 41-45. https://doi.org/10.33584/jnzg.2007.69.2684.

Pozza, L.E., Bishop, T.F.A., Stockmann, U., Birch, G.F., 2020. Integration of vis-NIR and pXRF spectroscopy for rapid measurement of soil lead concentrations. Soil Res. 58(3), 247-257. https://doi.org/10.1071/sr19174.

Prasad, A.S., 2012. Discovery of human zinc deficiency: 50 years later. J. Trace Elem. Med. Biol. 26(2), 66-69. https://doi.org/10.1016/j.jtemb.2012.04.004.

Pullanagari, R.R., Kereszturi, G., Yule, I.J., 2016. Mapping of macro and micro nutrients of mixed pastures using airborne AisaFENIX hyperspectral imagery. ISPRS J. Photogramm. Remote Sens. 117, 1-10. https://doi.org/10.1016/j.isprsjprs.2016.03.010.

Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H., Sorooshian, S., 1994. A modified soil adjusted vegetation index. Remote Sens. Environ. 48(2), 119-126. https://doi.org/10.1016/0034-4257(94)90134-1.

R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. R version 4.0.3. http://www.R-project.org/. .

Rajendram, G., Devey, K., 2011. Comparison between mid-infrared and near-infrared for soil analysis. In: L.D. Currie, C.L. Christensen (Eds.), Adding to the knoweldge base for the nutrient manager. FLRC, Massey University, Manawatu Campus, New Zealand.

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.

Ranga Suri, N., Murty M, N., Athithan, G., 2019. Outlier Detection: Techniques and Applications A Data Mining Perspective. Springer. https://doi.org/10.1007/978-3-030-05127-3.

Rathod, P.H., Brackhage, C., Muller, I., Van der Meer, F.D., Noomen, M.F., 2018. Assessing metal-induced changes in the visible and near-infrared spectral reflectance of leaves: A pot study with sunflower

(Helianthus annuus L.). J. Indian Soc. Remote Sens. 46(12), 1925-1937. https://doi.org/10.1007/s12524-018-0846-3.
Rathod, P.H., Brackhage, C., Van der Meer, F.D., Muller, I., Noomen, M.F., Rossiter, D.G., Dudel, G.E., 2015a. Spectral changes in the leaves of barley plant due to phytoremediation of metals - results from a pot study. Eur. J. Remote Sens. 48(1), 283-302. https://doi.org/10.5721/EuJRS20154816.
Rathod, P.H., Muller, I., Van der Meer, F.D., de Smeth, B., 2015b. Analysis of visible and near infrared spectral reflectance for assessing metals in soil. Environ. Monit. Assess. 188(10), 558. https://doi.org/10.1007/s10661-016-5568-9.
Rathod, P.H., Rossiter, D.G., Noomen, M.F., van der Meer, F.D., 2013. Proximal spectral sensing to monitor phytoremediation of metal-contaminated soils. 15(5), 405-426. https://doi.org/10.1080/15226514.2012.702805.
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(3), 347-357. https://doi.org/10.1080/00288233.2018.1426616.
Rattenbury, M., Martin, A.P., Turnbull, R., Christie, A., 2014. Sampling methodology for a regional multi-element geochemical baseline survey. GNS Science report 2014/62, GNS Science, New Zealand, Lower Hutt.
Ravansari, R., Lemke, L.D., 2018. Portable X-ray fluorescence trace metal measurement in organic rich soils: pXRF response as a function of organic matter fraction. Geoderma 319, 175-184. https://doi.org/10.1016/j.geoderma.2018.01.011.
Ravansari, R., Wilson, S.C., Tighe, M., 2020. Portable X-ray fluorescence for environmental assessment of soils: Not just a point and shoot method. Environ. Int. 134, 105250. https://doi.org/10.1016/j.envint.2019.105250.
Reeves III, J.B., Smith, D.B., 2009. The potential of mid-and near-infrared diffuse reflectance spectroscopy for determining major-and trace-element concentrations in soils from a geochemical survey of North America. Appl. Geochem. 24(8), 1472-1481. https://doi.org/10.1016/j.apgeochem.2009.04.017.
Reimann, C., de Caritat, P., 2012. New soil composition data for Europe and Australia: Demonstrating comparability, identifying continental-scale processes and learning lessons for global geochemical mapping. Sci. Total Environ. 416, 239-252. https://doi.org/10.1016/j.scitotenv.2011.11.019.
Reiser, R., Simmler, M., Portmann, D., Clucas, L., Schulin, R., Robinson, B., 2014. Cadmium concentrations in New Zealand pastures: relationships to soil and climate variables. J. Environ. Qual. 43(3), 917-925. https://doi.org/10.2134/jeq2013.09.0367.
Revelle, W., 2021. psych: procedures for pyschological, pyschometric and personality research, Northwestern University, Evanston, Illinois, USA, R package version 2.1.6.
Roberts, A.H.C., Longhurst, R.D., Brown, M.W., 1994. Cadmium status of soils, plants, and grazing animals in New Zealand. N. Z. J. Agric. Res. 37(1), 119-129. https://doi.org/10.1080/00288233.1994.9513048.
Roberts, J., Power, A., Chapman, J., Chandra, S., Cozzolino, D., 2018. Vibrational spectroscopy methods for agro-food product analysis, Comprehensive Analytical Chemistry. Elsevier B.V.
Rondeaux, G., Steven, M., Baret, F., 1996. Optimization of soil-adjusted vegetation indices. Remote Sens. Environ. 55(2), 95-107. https://doi.org/10.1016/0034-4257(95)00186-7.
Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158(1-2), 46-54. https://doi.org/10.1016/j.geoderma.2009.12.025.
Rossel, R.A.V., Jeon, Y.S., Odeh, I.O.A., McBratney, A.B., 2008. Using a legacy soil sample to develop a mid-IR spectral library. Aust. J. Soil Res. 46(1), 1-16. https://doi.org/10.1071/Sr07099.
Rossel, R.A.V., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. Eur. J. Soil Sci. 63(6), 848-860. https://doi.org/10.1111/j.1365-2389.2012.01495.x.
Rosso, P.H., Pushnik, J.C., Lay, M., Ustin, S.L., 2005. Reflectance properties and physiological responses of Salicornia virginica to heavy metal and petroleum contamination. Environ. Pollut. 137(2), 241-252. https://doi.org.10.1016/j.envpol.2005.02.025.
Roudier, P., 2021. spectacles: storing and manipulating spectroscopy data in R. R package version 0.5-3.
Roudier, P., Burge, O.R., Richardson, S.J., McCarthy, J.K., Grealish, Gerard J., Ausseil, A.-G., 2020. National scale 3D mapping of soil pH using a data augmentation approach. Remote Sens. 12(18), 2872. https://doi.org/10.3390/rs12182872.
Roudier, P., Hedley, C.B., Lobsey, C.R., Viscarra Rossel, R.A., Leroux, C., 2017. Evaluation of two methods to eliminate the effect of water from soil vis–NIR spectra for predictions of organic carbon. Geoderma 296, 98-107. https://doi.org/10.1016/j.geoderma.2017.02.014.

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.
Roujean, J.-L., Breon, F.-M., 1995. Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. Remote Sens. Environ. 51(3), 375-384. https://doi.org/10.1016/0034-4257(94)00114-3.
RStudio Team, 2021. RStudio: Integrated development environment for R. RStudio, PBC, Boston, MA. R Studio version: 1.4.1103. http://www.rstudio.com/. .
Rustam, Z., Sudarsono, E., Sarwinda, D., 2019. Random-forest (RF) and support vector machine (SVM) implementation for analysis of gene expression data in chronic kidney disease (CKD). IOP Conf. Ser.-Mat. Sci. 546, 052066. https://doi.org/10.1088/1757-899x/546/5/052066.
Rys, G., Newland, S., Keenan, C., Court, J., Furness, H., Sneath, G., Kim, N., Noiton, D., Taylor, T., Grogan, E., Manning, M., Roberts, A., Catto, W., Clear, M., Dansted, P., Young, B., Bedford, G., Hellstrom, J., Taylor, M., Faulkner, A., 2008. Cadmium in New Zealand Agriculture Report of the Cadmium Working Group.
Salmanzadeh, M., 2017. Cadmium accumulation in agricultural soils. Doctoral Thesis, The University of Waikato, Hamilton, New Zealand.
Salmanzadeh, M., Balks, M.R., Hartland, A., Schipper, L.A., 2016. Cadmium accumulation in three contrasting New Zealand soils with the same phosphate fertilizer history. Geoderma Reg. 7(3), 271-278. https://doi.org/10.1016/j.geodrs.2016.05.001.
Salminen, R., Plant, J.A., Reeder, S., 2005. Geochemical Atlas of Europe, Part 1, Background Information, Methodology and Maps. Geological Survey of Finland, Espoo.
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(2), 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(8), 1627-1639. https://doi.org/10.1021/ac60214a047.
Schneider, A.R., Cances, B., Breton, C., Ponthieu, M., Morvan, X., Conreux, A., Marin, B., 2016. Comparison of field portable XRF and aqua regia/ICPAES soil analysis and evaluation of soil moisture influence on FPXRF results. J. Soil Sediment 16(2), 438-448. https://doi.org/10.1007/s11368-015-1252-x.
Shamsoddini, A., Raval, S., Taplin, R., 2014. Spectroscopic analysis of soil metal contamination around a derelict mine site in the Blue mountains, Australia. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 2(7), 75-79. https://doi.org/10.5194/isprsannals-II-7-75-2014.
Sharma, P., Dubey, R., 2006. Cadmium uptake and its toxicity in higher plants. In: N. Khan, Samiullah (Eds.), Cadmium toxicity and tolerance in plants. Narosa Publishing House, New Delhi, pp. 63-86.
Shenk, J.S., Westerhaus, M.O., Berzaghi, P., 1997. Investigation of a LOCAL calibration procedure for near Infrared Instruments. J. Near Infrared Spectrosc. 5(4), 223-232. https://doi.org/10.1255/jnirs.115.
Shepherd, K.D., Ferguson, R., Hoover, D., Van Egmond, F., Sanderman, J., Ge, Y., 2022. A global soil spectral calibration library and estimation service. Soil Secur., 100061. https://doi.org/10.1016/j.soisec.2022.100061.
Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66(3), 988-998. https://doi.org/10.2136/sssaj2002.9880.
Shepherd, K.D., Walsh, M.G., 2007. Infrared spectroscopy - enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries. J. Near Infrared Spectrosc. 15(1), 1-19. https://doi.org/10.1255/jnirs.716.
Shi, T.Z., Wang, J.J., Chen, Y.Y., Wu, G.F., 2016. Improving the prediction of arsenic contents in agricultural soils by combining the reflectance spectroscopy of soils and rice plants. Int. J. Appl. Earth Obs. 52, 95-103. https://doi.org/10.1016/j.jag.2016.06.002.
Shi, Z., Ji, W., Viscarra Rossel, R.A., Chen, S., Zhou, Y., 2015. Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis–NIR spectral library. Eur. J. Soil Sci. 66(4), 679-687. https://doi.org/10.1111/ejss.12272.
Shin, J.H., Yu, J., Wang, L., Kim, J., Koh, S.M., Kim, S.O., 2019. Spectral responses of heavy metal contaminated soils in the vicinity of a hydrothermal ore deposit: A case study of Boksu mine,

South Korea. IEEE Trans. Geosci. Remote Sens. 57(6), 4092-4106. https://doi.org/10.1109/Tgrs.2018.2889748.
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, S., Becker, M., Lamers, J.P.A., Wimmer, M.A., 2021. Residual effects of B and Zn fertilizers applied to dry season crops on the performance of the follow-up crop of maize in Nepal. J. Plant Nutr. Soil Sci. 184(2), 238-245. https://doi.org/10.1002/jpln.202000289.
Shuman, L.M., 1985. Fractionation method for soil microelements. Soil Sci. 140(1), 11-22. https://doi.org/10.1097/00010694-198507000-00003.
Siebers, N., Kruse, J., Leinweber, P., 2013. Speciation of phosphorus and cadmium in a contaminated soil amended with bone char: Sequential fractionations and XANES spectroscopy. Water Air Soil Pollut. 224(5), 1564. https://doi.org/10.1007/s11270-013-1564-7.
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(6), 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. Chemometr. Intell. Lab. Syst. 153, 92-105. https://doi.org/10.1016/j.chemolab.2016.02.013.
Smith, R., Adams, J., Stephens, D., Hick, P., 1995. Forecasting wheat yield in a Mediterranean-type environment from the NOAA satellite. Aus. J. Agric. Res. 46(1), 113. https://doi.org/10.1071/ar9950113.
Snee, R.D., 1977. Validation of regression models: methods and examples. Technomet. 19(4), 415-428. https://doi.org/10.1080/00401706.1977.10489581.
Soil Bureau, 1962. New Zealand Soil Bureau maps. Department of Scientific and Industrial Research, Wellington, New Zealand.
Song, Y., Li, F., Yang, Z., Ayoko, G.A., Frost, R.L., Ji, J., 2012. Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang River Delta, China. Appl. Clay Sci. 64, 75-83. https://doi.org/10.1016/j.clay.2011.09.010.
Soriano-Disla, J.M., Janik, L., McLaughlin, M.J., Forrester, S., Kirby, J., Reimann, C., Team, E.G.P., 2013. The use of diffuse reflectance mid-infrared spectroscopy for the prediction of the concentration of chemical elements estimated by X-ray fluorescence in agricultural and grazing European soils. Appl. Geochem. 29, 135-143. https://doi.org/0.1016/j.apgeochem.2012.11.005.
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(2), 139-186. https://doi.org/10.1080/05704928.2013.811081.
Spark, K.M., Johnson, B.B., Wells, J.D., 1995. Characterizing heavy-metal adsorption on oxides and oxyhydroxides. Eur. J. Soil Sci. 46(4), 621-631. https://doi.org/10.1111/j.1365-2389.1995.tb01358.x.
Sposito, G., 1984. The surface chemistry of soils. Oxford university press.
Sposito, G., Lund, L.J., Chang, A.C., 1982. Trace metal chemistry in arid-zone field soils amended with sewage sludge: I. Fractionation of Ni, Cu, Zn, Cd, and Pb in solid phases. Soil Sci. Soc. Am. J. 46(2), 260-264. https://doi.org/10.2136/sssaj1982.03615995004600020009x.
Sridhar, B.B.M., Han, F.X., Diehl, S.V., Monts, D.L., Su, Y., 2007. Spectral reflectance and leaf internal structure changes of barley plants due to phytoextraction of zinc and cadmium. Int. J. Remote Sens. 28(5), 1041-1054. https://doi.org/10.1080/01431160500075832.
Stafford, A., Jeyakumar, P., Hedley, M., Anderson, C., 2018a. Influence of soil moisture status on soil cadmium phytoavailability and accumulation in plantain (Plantago lanceolata). Soil Syst. 2(1). https://doi.org/10.3390/soils2010009.
Stafford, A., Kusumo, B., Jeyakumar, P., Hedley, M., Anderson, C., 2018b. Cadmium in soils under pasture predicted by soil spectral reflectance on two dairy farms in New Zealand. Geoderma Reg. 13, 26-34. https://doi.org/10.1016/j.geodrs.2018.03.001.
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. Doctoral Thesis, Massey University, Manawatu Campus, New Zealand, 324 pp. https://mro.massey.ac.nz/handle/10179/12980.
Stafford, A.D., Anderson, C.W.N., Hedley, M.J., McDowell, R.W., 2016. Cadmium accumulation by forage species used in New Zealand livestock grazing systems. Geoderma Reg. 7(1), 11-18. https://doi.org/10.1016/j.geodrs.2015.11.003.

Stafford, A.D., Palmer, A.S., Jeyakumar, P., Hedley, M.J., Anderson, C.W.N., 2018c. Soil cadmium and New Zealand dairy farms: Impact of whole-farm contaminant variability on environmental management. Agric. Ecosyst. Environ. 254, 282-291. https://doi.org/10.1016/j.agee.2017.11.033.
Stefansson, P., Liland, K.H., Thiis, T., Burud, I., 2020. Fast method for GA-PLS with simultaneous feature selection and identification of optimal preprocessing technique for datasets with many observations. J. Chemometr. 34(3), e3195. https://doi.org/10.1002/cem.3195.
Stenberg, B., Rossel, R.A.V., 2010. Diffuse reflectance spectroscopy for high-resolution soil sensing. In: R.A. Viscarra Rossel, A.B. McBratney, B. Minasny (Eds.), Proximal Soil Sensing. Springer Netherlands, Dordrecht, pp. 29-47. https://doi.org/10.1007/978-90-481-8859-8_3.
Stenberg, B., Viscarra Rossel, R., Mouazen, A., Wetterlind, J., 2010. Chapter Five - Visible and Near Infrared Spectroscopy in Soil Science. Adv. Agron. 107, 163-215. https://doi.org/10.1016/S0065-2113(10)07005-7.
Stevens, A., Ramirez Lopez, L., 2020. An introduction to prospectr. R package version 0.2.1.
Stojsavljevic, A., Rovcanin, B., Jagodic, J., Krstic, D., Paunovic, I., Gavrovic-Jankulovic, M., Manojlovic, D., 2021. Alteration of trace elements in multinodular goiter, thyroid adenoma, and thyroid cancer. Biol. Trace Elem. Res. https://doi.org/10.1007/s12011-020-02542-9.
Stosnach, H., 2005. Environmental trace-element analysis using a benchtop total reflection X-ray fluorescence spectrometer. Anal. Sci. 21(7), 873-876. https://doi.org/10.2116/analsci.21.873.
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(2), 693-715. https://doi.org/10.5194/soil-7-693-2021.
Tay, S.L., Scott, J.M., Craw, D., 2021. Natural rehabilitation of arsenic-rich historical tailings at the Alexander mine, Reefton, New Zealand. N. Z. J. Geol. Geophys. 64(4), 558-569. https://doi.org/10.1080/00288306.2020.1840402.
Taylor, M., Gibb, R., Willoughby, J., Hewitt, A., Arnold, G., 2007. Soil maps of cadmium in New Zealand, Ministry of Agriculture and Forestry, Ltd., Manaaki Whenua-Landcare Research, Lincoln, New Zealand.
Taylor, M.D., Kim, N.D., Hill, R.B., Chapman, R., 2010. A review of soil quality indicators and five key issues after 12 yr soil quality monitoring in the Waikato region. Soil Use Manag. 26(3), 212-224. https://doi.org/10.1111/j.1475-2743.2010.00276.x.
Temminghoff, E.J.M., Zee, S.E.A.T.M., Haan, F.A.M., 1995. Speciation and calcium competition effects on cadmium sorption by sandy soil at various pHs. Eur. J. Soil Sci. 46(4), 649-655. https://doi.org/10.1111/j.1365-2389.1995.tb01361.x.
Tessier, A., Campbell, P.G.C., Bisson, M., 1979. Sequential extraction procedure for the speciation of particulate trace-metals. Anal. Chem. 51(7), 844-851. https://doi.org/10.1021/ac50043a017.
Thomas, C.L., Hernandez-Allica, J., Dunham, S.J., McGrath, S.P., Haefele, S.M., 2021. A comparison of soil texture measurements using mid-infrared spectroscopy (MIRS) and laser diffraction analysis (LDA) in diverse soils. Sci. Rep. 11(1), 16. https://doi.org/10.1038/s41598-020-79618-y.
Thompson, B., 2019. A limitation of random forest regression.
Thompson-Morrison, H., 2017. Cadmium management in New Zealand's horticultural soils. Master degree Thesis, Massey University, Palmerston North, New Zealand.
Todorova, M., Mouazen, A.M., Lange, H., Atanassova, S., 2014. Potential of near-infrared spectroscopy for measurement of heavy metals in soil as affected by calibration set size. Water Air Soil Pollut. 225(8), 1. https://doi.org/10.1007/s11270-014-2036-4.
Towett, E.K., Shepherd, K.D., Lee Drake, B., 2015a. Plant elemental composition and portable X-ray fluorescence (pXRF) spectroscopy: quantification under different analytical parameters. X-Ray Spectrom. 45(2), 117-124. https://doi.org/10.1002/xrs.2678.
Towett, E.K., Shepherd, K.D., Sila, A., Aynekulu, E., Cadisch, G., 2015b. Mid-infrared and total X-ray fluorescence spectroscopy complementarity for assessment of soil properties. Soil Sci. Soc. Am. J. 79(5), 1375-1385. https://doi.org/10.2136/sssaj2014.11.0458.
Traina, S., 1999. The environmental chemistry of cadmium. In: M.J. McLaughlin, B.R. Singh (Eds.), Cadmium in soils and plants. Developments in Plant and Soil Sciences. Springer, Dordrecht, NL., pp. 11-37. https://doi.org/10.1007/978-94-011-4473-5_2.
Trucano, T.G., Swiler, L.P., Igusa, T., Oberkampf, W.L., Pilch, M., 2006. Calibration, validation, and sensitivity analysis: what's what. Reliab. Eng. Syst. Saf. 91(10-11), 1331-1357. https://doi.org/10.1016/j.ress.2005.11.031.

Turnbull, R., Rogers, K., Martin, A., Rattenbury, M., Morgan, R., 2019. Human impacts recorded in chemical and isotopic fingerprints of soils from Dunedin City, New Zealand. Sci. Total Environ. 673, 455-469. https://doi.org/10.1016/j.scitotenv.2019.04.063.

UAE, 2020. Infrared spectrometers: NIR and MIR compared. University of East Anglia, United Kingdom.

Ubeynarayana, N., Jeyakumar, P., Bishop, P., Pereira, R.C., Anderson, C.W.N., 2021. Effect of soil cadmium on root organic acid secretion by forage crops. Environ. Pollut. 268(Pt A), 115839. https://doi.org/10.1016/j.envpol.2020.115839.

Usman, K., Al-Ghouti, M.A., Abu-Dieyeh, M.H., 2019. The assessment of cadmium, chromium, copper, and nickel tolerance and bioaccumulation by shrub plant Tetraena qataranse. Sci. Rep. 9(1), 5658. https://doi.org/10.1038/s41598-019-42029-9.

Vasques, G.M., Dematte, J.A.M., Rossel, R.A.V., Ramirez-Lopez, L., Terra, F.S., 2014. Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. Geoderma 223, 73-78. https://doi.org/10.1016/j.geoderma.2014.01.019.

Viscarra Rossel, R.A., Adamchuk, V.I., Sudduth, K.A., McKenzie, N.J., Lobsey, C., 2011. Proximal soil sensing: An effective approach for soil measurements in space and time. In: D.L. Sparks (Ed.), Advances in Agronomy. Academic Press, pp. 243-291. https://doi.org/10.1016/b978-0-12-386473-4.00005-1.

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., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131(1-2), 59-75. https://doi.org/10.1016/j.geoderma.2005.03.007.

Vogelmann, J.E., Rock, B.N., Moss, D.M., 1993. Red edge spectral measurements from sugar maple leaves. Int. J. Remote Sens. 14(8), 1563-1575. https://doi.org/10.1080/01431169308953986.

Wadoux, A., Malone, B., Minasny, B., Fajardo, M., McBratney, A.B., 2021a. Soil Spectral Inference with R. Springer. https://doi.org/10.1007/978-3-030-64896-1.

Wadoux, A.M.J.C., Román-Dobarco, M., McBratney, A.B., 2021b. Perspectives on data-driven soil research. Eur. J. Soil Sci. 72(4), 1675-1689. https://doi.org/10.1111/ejss.13071.

Waikato Regional Council, 2016. Regional policy statement for the Waikato Region, Hamilton, NZ, Waikato Regional Council.

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), 2045-2322. https://doi.org/10.1038/srep40709.

Wang, D.D., Chakraborty, S., Weindorf, D.C., Li, B., Sharma, A., Paul, S., Ali, M.N., 2015. Synthesized use of VisNIR DRS and PXRF for soil characterization: Total carbon and total nitrogen. Geoderma 243, 157-167. https://doi.org/10.1016/j.geoderma.2014.12.011.

Wang, F.H., Gao, J., Zha, Y., 2018. Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges. ISPRS J. Photogramm. Remote Sens. 136, 73-84. https://doi.org/10.1016/j.isprsjprs.2017.12.003.

Wang, J.J., Cui, L.J., Gao, W.X., Shi, T.Z., Chen, Y.Y., Gao, Y., 2014. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. Geoderma 216, 1-9. https://doi.org/10.1016/j.geoderma.2013.10.024.

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. https://doi.org/10.1016/j.rse.2022.112914.

Wang, Z., Zeng, X., Yu, X., Zhang, H., Li, Z., Jin, D., 2010. Adsorption behaviors of Cd2+ on Fe2O3/MnO2 and the effects of coexisting ions under alkaline conditions. Chin. J. Geochem. 29(2), 197-203. https://doi.org/10.1007/s11631-009-0197-4.

Weindorf, D.C., Chakraborty, S., 2020. Portable X-ray fluorescence spectrometry analysis of soils. Soil Sci. Soc. Am. J. 84(5), 1384-1392. https://doi.org/10.1002/saj2.20151.

Weindorf, D.C., Zhu, Y., Chakraborty, S., Bakr, N., Huang, B., 2012. Use of portable X-ray fluorescence spectrometry for environmental quality assessment of peri-urban agriculture. Environ. Monit. Assess. 184(1), 217-227. https://doi.org/10.1007/s10661-011-1961-6.

Weiss, C., Roetzer, G., 2016. GeomComb: (Geometric) Forecast Combination Methods for R. R package version 1.

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(6), 823-843. https://doi.org/10.1111/j.1365-2389.2010.01283.x.

White, K.E., Reeves, J.B., Coale, F.J., 2011. Mid-infrared diffuse reflectance spectroscopy for the rapid analysis of plant root composition. Geoderma 167-168, 197-203. https://doi.org/10.1016/j.geoderma.2011.08.009.

Whitehead, D., McNeill, S.J.E., Mudge, P.L., 2021. Regional and national changes in soil carbon stocks with land-use change from 1990 to 2016 for New Zealand. 21(4). https://doi.org/10.1007/s10113-021-01837-4.

Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration problem in chemistry solved by the PLS method. In: A. Ruhe, B. Kagstrom (Eds.), Proceeding Conference Matrix Pencils, Lecture Notes in Mathmatics. Springer-Verlag, Heidelberg, pp. 286-293. http://dx.doi.org/10.1007/BFB0062108.

Wu, C., Liu, M., Liu, X., Wang, T., Wang, L., 2019. Developing a new spectral index for detecting cadmium-induced stress in rice on a regional scale. Int. J. Environ. Res. Public Health 16(23). https://doi.org/10.3390/ijerph16234811.

Wu, C.-M., Tsai, H.-T., Yang, K.-H., Wen, J.-C., 2012. How reliable is X-ray fluorescence (XRF) measurement for different metals in soil contamination? Environ. Forensics. 13(2), 110-121. https://doi.org/10.1080/15275922.2012.676603.

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(1-4), 377-390. https://doi.org/10.1007/s11270-009-0206-6.

Wu, D., He, Y., Shi, J., Feng, S., 2009. Exploring near and midinfrared spectroscopy to predict trace iron and zinc contents in powdered milk. J. Agric. Food Chem. 57(5), 1697-1704. https://doi.org/10.1021/jf8030343.

Wu, Y.Z., Chen, J., Ji, J.F., Gong, P., Liao, Q.L., Tian, Q.J., Ma, H.R., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. Soil Sci. Soc. Am. J. 71(3), 918-926. https://doi.org/10.2136/sssaj2006.0285.

Xia, X.Q., Mao, Y.Q., Ji, J.F., Rui, H., Chen, J., Liao, Q.L., 2007. Reflectance spectroscopy study of Cd contamination in the sediments of the Changjiang river, China. Environ. Sci. Technol. 41(10), 3449-3454. https://doi.org/10.1021/es0624422.

Xian, X., Shokohifard, G.I., 1989. Effect of pH on chemical forms and plant availability of cadmium, zinc, and lead in polluted soils. Water Air Soil Pollut. 45(3-4), 265-273. https://doi.org/10.1007/BF00283457.

Xie, X.-L., Pan, X.-Z., Sun, B., 2012. Visible and near-infrared diffuse reflectance spectroscopy for prediction of soil properties near a copper smelter. Pedosphere 22(3), 351-366. https://doi.org/10.1016/s1002-0160(12)60022-8.

Xu, D., Chen, S., Xu, H., Wang, N., Zhou, Y., Shi, Z., 2020. Data fusion for the measurement of potentially toxic elements in soil using portable spectrometers. Environ. Pollut. 263(Pt A), 114649. https://doi.org/10.1016/j.envpol.2020.114649.

Xu, D., Zhao, R., Li, S., Chen, S., Jiang, Q., Zhou, L., Shi, Z., 2019a. Multi-sensor fusion for the determination of several soil properties in the Yangtze river delta, China. Eur. J. Soil Sci. 70(1), 162-173. https://doi.org/10.1111/ejss.12729.

Xu, D.Y., Chen, S.C., Rossel, R.A.V., Biswas, A., Li, S., Zhou, Y., Shi, Z., 2019b. X-ray fluorescence and visible near infrared sensor fusion for predicting soil chromium content. Geoderma 352, 61-69. https://doi.org/10.1016/j.geoderma.2019.05.036.

Xue, J., Su, B., 2017. Significant remote sensing vegetation indices: A review of developments and applications. J. Sensors 2017, 1353691. https://doi.org/10.1155/2017/1353691.

Yang, J., Hang, Z., Jiang, N., Li, Z., Bonheur, G., Zu, Y., 2021. Comparison of cadmium and arsenic accumulation characteristics and remediation potential of different forage plants, IOP Conference Series: Earth and Environmental Science. IOP Publishing, pp. 012057. https://doi.org/10.1088/1755-1315/787/1/012057.

Yi, Z., Lehto, N.J., Robinson, B.H., Cavanagh, J.E., 2020. Environmental and edaphic factors affecting soil cadmium uptake by spinach, potatoes, onion and wheat. Sci. Total Environ. 713, 136694. https://doi.org/10.1016/j.scitotenv.2020.136694.

Young, S.D., 2013. Chemistry of heavy metals and metalloids in soils. In: B.J. Alloway (Ed.), Heavy metals in soils: trace metals and metalloids in soils. Springer, Dordrecht, Netherlands, pp. 51-96. https://doi.org/10.1007/978-94-007-4470-7_3.

Yuan, G., Lavkulich, L.M., 1997. Sorption behavior of copper, zinc, and cadmium in response to simulated changes in soil properties. Commun. Soil Sci. Plant. 28(6-8), 571-587. https://doi.org/10.1080/00103629709369812.

Zanders, J.M., 1998. Studies on the origin, distribution and mobility of cadmium in pastoral soils. Doctoral Thesis, Massey University, New Zealand, 289 pp. https://mro.massey.ac.nz/bitstream/handle/10179/2420/02_whole.pdf.

Zanders, J.M., Hedley, M.J., Palmer, A.S., Tillman, R.W., Lee, J., 1999. The source and distribution of cadmium in soils on a regularly fertilised hill-country farm. Aust. J. Soil Res. 37(4), 667-678. https://doi.org/10.1071/SR98080.

Zarco-Tejada, P.J., Berjón, A., López-Lozano, R., Miller, J.R., Martín, P., Cachorro, V., González, M.R., de Frutos, A., 2005. Assessing vineyard condition with hyperspectral indices: Leaf and canopy reflectance simulation in a row-structured discontinuous canopy. Remote Sens. Environ. 99(3), 271-287. https://doi.org/10.1016/j.rse.2005.09.002.

Zarco-Tejada, P.J., Miller, J.R., Noland, T.L., Mohammed, G.H., Sampson, P.H., 2001. Scaling-up and model inversion methods with narrowband optical indices for chlorophyll content estimation in closed forest canopies with hyperspectral data. IEEE Transac. Geosci. Remote Sens. 39(7), 1491-1507. https://doi.org/10.1109/36.934080.

Zayed, A., Gowthaman, S., Terry, N., 1998. Phytoaccumulation of trace elements by wetland plants: I. Duckweed. J. Environ. Qual. 27(3), 715-721. https://doi.org/10.2134/jeq1998.00472425002700030032x.

Zhang, C., Feng, X., Wang, J., Liu, F., He, Y., Zhou, W., 2017. Mid-infrared spectroscopy combined with chemometrics to detect Sclerotinia stem rot on oilseed rape (Brassica napus L.) leaves. Plant Methods 13(1), 39. https://doi.org/10.1186/s13007-017-0190-6.

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(Pt 1), 321-334. https://doi.org/10.1016/j.scitotenv.2018.08.442.

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. https://doi.org/10.1111/ejss.12875.

Zimmermann, M., Leifeld, J., Fuhrer, J., 2007. Quantifying soil organic carbon fractions by infrared-spectroscopy. Soil Biol. Biochem. 39(1), 224-231. https://doi.org/10.1016/j.soilbio.2006.07.010.

Appendices

Appendix 2.1 Wavelengths assignments for overtones and combinations of fundamental molecular vibrations of soil constituents in the vis-NIR regions.

Wavelengths (nm)	Soil constituents	References
400	Ferric	(Hunt and Slisbury, 1970)
400	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
404	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
425	Iron	(Grove et al., 1992), (Ben-Dor and Banin, 1995), (Ben-Dor and Banin, 1990a)
430	Ferrous	(Hunt and Slisbury, 1970)
434	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
444	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
450	Ferrous	(Hunt and Slisbury, 1970)
concave between 450–850	Iron oxides	(Demattê et al., 2004)
450	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
500	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
510	Ferrous	(Hunt and Slisbury, 1970)
529	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
550	Ferrous	(Hunt and Slisbury, 1970)
564.4	Organic matter	(Krishnan et al., 1980)
623.6	Organic matter	(Krishnan et al., 1980)
640	Iron	(Ben-Dor and Banin, 1990a), (Grove et al., 1992), (Ben-Dor and Banin, 1995)
650	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
650	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
700	Ferric	(Hunt and Slisbury, 1970)
870	Ferric	(Hunt and Slisbury, 1970)
884	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
900	Ferric	(Demattê et al., 2004)
900	Iron	(Grove et al., 1992), (Ben-Dor and Banin, 1995), (Ben-Dor and Banin, 1990a)
900	Organic C	Beck et al 1976
900	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
920	Iron oxides	(Rossel and Behrens, 2010), (Soriano-Disla et al., 2014)
1000	Ferrous	(Hunt and Slisbury, 1970)
1000	Iron	(Hunt et al., 1974)
1025	Iron	(Grove et al., 1992), (Ben-Dor and Banin, 1995), (Ben-Dor and Banin, 1990a)

Wavelengths (nm)	Soil constituents	References
1075	Iron	(Grove et al., 1992), (Ben-Dor and Banin, 1995), (Ben-Dor and Banin, 1990a)
1100	Organic matter	(Ben-Dor and Banin, 1995)
1220	Organic C	Beck et al 1976
1266.1	Iron	(Ben-Dor and Banin, 1990a), (Grove et al., 1992), (Ben-Dor and Banin, 1995)
1300	Smectite	(Ben-Dor, 2002)
1300–1450	Active spectral region for Al containing smectites (2:1 configuration)	(Ben-Dor, 2002)
1400	Kaolinite, gibbsite	(Grove et al., 1992)
1400	Smectite	(Cariati, 1983)
1400	Kaolinite	(Rossel and Behrens, 2010)
1395–1415	Overtones of O–H stretch	(Stenberg et al., 2010)
1450	Water	(Hunt and Slisbury, 1970)
1456	Ca-Montmorillonite	(Hunt and Slisbury, 1970), (Ben-Dor and Banin, 1990a)
1600	Organic matter	(Ben-Dor and Banin, 1995)
1700	Organic matter	(Ben-Dor and Banin, 1995)
1726	Organic matter	(Morra et al., 1991)
1800	Carbonates	(Ben-Dor and Banin, 1990b), (Soriano-Disla et al., 2014)
1800–1900	Smectite	(Ben-Dor, 2002)
1800	Organic matter	(Ben-Dor and Banin, 1995)
1900	Clay mineral Strong (2:1) Weak (1:1)	(Demattê et al., 2004)
1900	Smectite	(Cariati, 1983)
1910	Ca-Montmorillonite	(Hunt and Slisbury, 1970), (Ben-Dor and Banin, 1990a)
1950	Water	(Hunt and Slisbury, 1970)
1978	Ca-Montmorillonite	(Hunt and Slisbury, 1970), (Ben-Dor and Banin, 1990a)
2000	Organic matter	(Ben-Dor and Banin, 1995)
2160	Al–OH	(Rossel and Behrens, 2010)
2200	Kaolinite	(Demattê et al., 2004)
2165–2207	Al–OH bend plus O-H stretch combination	(Stenberg et al., 2010)
2200	Absorption due to Al–OH	(Stenberg et al., 2010)
2200–2500	Smectite	(Ben-Dor, 2002)
2200–2500	Combination vibrations involving a O-H stretch and metal–OH bend	(Stenberg et al., 2010)
2200	Smectite	(Cariati, 1983)
2200	Smectite	(Rossel and Behrens, 2010)
2200–2400	Organic matter	(Ben-Dor and Banin, 1995)
2200	Illite	(Rossel and Behrens, 2010)
2200	Kaolinite	(Rossel and Behrens, 2010)
2208	Al–OH	(Rossel and Behrens, 2010)
2230	Al–OH	(Rossel and Behrens, 2010)
2259.8	Iron	(Ben-Dor and Banin, 1990a), (Grove et al., 1992), (Ben-Dor and Banin, 1995)
2265	Gibbsite	(Demattê et al., 2004)
2294.9	Iron	(Ben-Dor and Banin, 1990a), (Grove et al., 1992), (Ben-Dor and Banin, 1995)
2335	Carbonates	(Ben-Dor and Banin, 1990b), (Soriano-Disla et al., 2014)
2340	Illite	(Rossel and Behrens, 2010)
2350	Carbonates	(Ben-Dor and Banin, 1990b), (Soriano-Disla et al., 2014)

Wavelengths (nm)	Soil constituents	References
2360	Carbonates	(Ben-Dor and Banin, 1990b), (Soriano-Disla et al., 2014)
2426	Organic matter	(Morra et al., 1991)
2445	Illite	(Rossel and Behrens, 2010)

Appendix 2.2 Vegetation indices developed using different wavelengths in the vis-NIR region.

Vegetation indices	Formulas	References
Anthocyanin reflectance index 1	$((1/R_{550}) - (1/R_{700}))$	(Gitelson et al., 2001)
Anthocyanin reflectance index 2	$R_{800} \times ((1/R_{550}) - (1/R_{700}))$	(Gitelson et al., 2001)
Blue green index 1	R_{400}/R_{550}	(Zarco-Tejada et al., 2005)
Blue green index 2	R_{450}/R_{550}	(Zarco-Tejada et al., 2005)
Blue red index 1	R_{400}/R_{690}	(Zarco-Tejada et al., 2005)
Blue red index 2	R_{450}/R_{690}	(Zarco-Tejada et al., 2005)
Cellulose absorption index	$0.5 \times (R_{2000} + R_{2200}) - R_{2100}$	(Daughtry, 2001)
Chlorophyll absorption ratio index	$((((R_{700} - R_{500})/R_{150}) \times R_{670} + R_{670} + (R_{550} - (((R_{700} - R_{500})/R_{150}) \times R_{550})))/ (((R_{700} - R_{500})/R_{150})^2 + 1)^0.5) \times (R_{700}/R_{670})$	(Daughtry, 2001)
Chlorophyll red index 500	$(1/R_{515})/(1/R_{550})$	(Gitelson et al., 2006)
Chlorophyll red index 700	$(1/R_{515})/(1/R_{700})$	(Gitelson et al., 2006)
Different vegetation index	$R_{780} - R_{670}$	(Buschmann and Nagel, 1993)
Greenness Index	R_{554}/R_{677}	(Smith et al., 1995)
Gitelson and Merzlyak 1	R_{750}/R_{550}	(Gitelson and Merzlyak, 1997)
Gitelson and Merzlyak 2	R_{750}/R_{700}	(Gitelson and Merzlyak, 1997)
Healthy index	$((R_{534} - R_{698})/(R_{534} - R_{698})) - 0.5 \times R_{704}$	(Mahlein et al., 2013)
Infrared percentage vegetation index	$R_{780}/(R_{780} + R_{670})$	(Crippen, 1990)
Lichtenthaler Indices 3	R_{400}/R_{740}	(Lichtenthaler et al., 1996)
Modified chlorophyll absorption ratio index 1	$((R_{700} - R_{670}) - 0.2 \times (R_{700} - R_{550})) \times (R_{700}/R_{670})$	(Daughtry et al., 2000)
Modified chlorophyll absorption ratio index 2	$(1.5 \times (2.5 \times (R_{800} - R_{670}) - 1.3 \times (R_{800} - R_{550})))/sqrt((2 \times R_{800} + 1)^2 - (6 \times R_{800} - 5 \times R_{670}) - 0.5)$	(Haboudane et al., 2004)
Modified soil adjusted vegetation index	$0.5 \times (2 \times R_{800} + 1 - sqrt((2 \times R_{800} + 1)^2 - 8 \times (R_{800} - R_{670})))$	(Qi et al., 1994)
Modified simple ratio	$((R_{800}/R_{670}) - 1) / sqrt(R_{800}/R_{670} + 1)$	(Chen, 1996)
Modified triangular vegetation index 1	$1.2 \times (1.2 \times (R_{800} - R_{550}) - 2.5 \times (R_{670} - R_{550}))$	(Haboudane et al., 2004)
Modified triangular vegetation index 2	$(1.5 \times (1.2 \times (R_{800} - R_{550}) - 2.5 \times (R_{670} - R_{550})))/sqrt((2 \times R_{800} + 1)^2 - (6 \times R_{800} - 5 \times sqrt(R_{670})) - 0.5)$	(Haboudane et al., 2004)
Normalised difference index 1	$(R_{780} - R_{710})/(R_{780} - R_{680})$	(Datt, 1999)
Normalised difference index 2	$(R_{850} - R_{710})/(R_{850} - R_{680})$	(Datt, 1999)
Normalised difference index 3	$(R_{734} - R_{747})/(R_{715} - R_{726})$	(Vogelmann et al., 1993)
Normalised difference vegetation index	$(R_{800} - R_{670})/(R_{800} + R_{670})$	(Lichtenthaler et al., 1996)
Normalised pigment chlorophyll index	$(R_{680} - R_{430})/(R_{680} + R_{430})$	(Peñuelas et al., 1994)

Vegetation indices	Formulas	References
Optimised soil adjusted vegetation index	$((1+ 0.16)/(R_{800} – R_{670}))/(R_{800} + R_{670} + 0.61))$	(Rondeaux et al., 1996)
Photochemical reflectance index	$(R_{531} – R_{570})/(R_{531} + R_{570})$	(Gamon et al., 1992)
Plant senescence reflectance spectra	$(R_{680} – R_{500})/R_{750}$	(Merzlyak et al., 1999)
Pigment specific normalised difference c	$(R_{800} – R_{470})/(R_{800} + R_{470})$	(Blackburn, 1998)
Pigment specific simple ratio a	R_{800}/R_{680}	(Blackburn, 1998)
Pigment specific simple ratio b	R_{800}/R_{635}	(Blackburn, 1998)
Pigment specific simple ratio c	R_{800}/R_{470}	(Blackburn, 1998)
Ratio analysis of reflectance spectra	R_{746}/R_{513}	(Gitelson and Merzlyak, 1997)
Renormalised difference vegetation index	$(R_{780} – R_{670})/(R_{780} + R_{670})$	(Roujean and Breon, 1995)
Red edge normalised difference vegetation index	$(R_{750} – R_{705})/(R_{750} + R_{705})$	(Gitelson and Merzlyak, 1994)
Red green reflectance ratio	sum $(R_{600}:R_{699}))$/sum $((R_{500}:R_{599}))$	(Gamon and Surfus, 1999)
Soil adjusted vegetation index	$((1+0.5) (R_{780} – R_{670}))/(R_{780} + R_{670} + 0.5)$	(Huete, 1988)
Simple ratio index	R_{800}/R_{670}	(Jordan, 1969)
Simple ratio index 2	R_{800}/R_{550}	(Buschmann and Nagel, 1993)
Simple ratio index 3	R_{700}/R_{670}	(McMurtrey et al., 1994)
Simple ratio index 4	R_{740}/R_{720}	(Vogelmann et al., 1993)
Simple ratio index 5	$R_{675}/(R_{700} \times R_{650})$	(Chappelle et al., 1992)
Simple ratio index 6	$R_{672}/(R_{550} \times R_{708})$	(Datt, 1998)
Simple ratio index 7	$R_{860}/(R_{550} \times R_{708})$	(Datt, 1998)
Simple ratio pigment index	R_{430}/R_{680}	(Peñuelas et al., 1994)
Structure insensitive pigment index	$(R_{800} – R_{445})/(R_{800} + R_{680})$	(PenUelas et al., 1995)
Transformed chlorophyll absorption in reflectance	$3 \times ((R_{700} – R_{670}) – 0.2 \times ((R_{700} – R_{550}) \times (R_{700} / R_{670})))$	(Haboudane et al., 2002)
Vogelmann red edge index 1	R_{740}/R_{720}	(Vogelmann et al., 1993)
Vogelmann red edge index 2	$(R_{734} – R_{747})/(R_{715} + R_{726})$	(Vogelmann et al., 1993)
Zarco and Miller index	R_{750}/R_{710}	(Zarco-Tejada et al., 2001)

Appendix 2.3 Wavenumbers assignments for fundamental MIR absorptions of soil constituents.

Wavenumbers (cm^{-1})	Soil constituents	References
3800	Kaolinite	(Stenberg et al., 2010)
3736	O–H, carbohydrate	(Wang et al., 2017)
3700–3000	Negative peak O–H of mineral compounds and H_2O)	(Zimmermann et al., 2007)
3705	Kaolinite, negative	(Janik et al., 1998)
3695	Kaolinite (O–H)	(Rossel and Behrens, 2010)
3695	Kaolinite, organic matter	(Niazi et al., 2015)
3690–3620	Kaolinite	(Soriano-Disla et al., 2014)
3657	Kaolinite, organic matter	(Niazi et al., 2015)
3631–3000	Smectite and gibbsite	(Janik et al., 1998)
3626	Kaolinite, organic matter	(Niazi et al., 2015)
3620	Clay minerals	(Rossel and Behrens, 2010)
3620	Smectite clay	(Janik et al., 1998)
3620–3630	Smectite and Illite	(Soriano-Disla et al., 2014)
3599	Kaolinite	(Stenberg et al., 2010)
3575	Hydroxyl (O–H)	(Rossel and Behrens, 2010)
3528	Gibbsite	(Niazi et al., 2015)
3484	Water	(Rossel and Behrens, 2010)
3461	Gibbsite	(Niazi et al., 2015)
3461	Kaolinite, organic matter	(Niazi et al., 2015)
3445	Smectite clay	(Janik et al., 1998)
3400–3300	Smectite and Illite	(Soriano-Disla et al., 2014)
3394	Hematite	(Niazi et al., 2015)
3394	Kaolinite, organic matter	(Niazi et al., 2015)
3330	Amine	(Rossel and Behrens, 2010)
3278	Water	(Rossel and Behrens, 2010)
3142	Goethite (–OH)	(Niazi et al., 2015)
3100	FeOOH	(Soriano-Disla et al., 2014)
3030	Aromatics	(Rossel and Behrens, 2010)
3000–2800	C-H stretch	(Zimmermann et al., 2007)
2930	Alkyl	(Rossel and Behrens, 2010)
2930–2850	Alkyl	(Soriano-Disla et al., 2014)
2929	Organic matter	(Stenberg et al., 2010)
2929–2855	Organic matter	(Janik et al., 1998)
2929–2855	Alkyl	(Janik et al., 1998)
2926–2850	carboxylates	(Janik et al., 1998)
2924	C-H	(Wang et al., 2017)
2850	Alkyl	(Rossel and Behrens, 2010)
2850	Organic matter	(Stenberg et al., 2010)
2520	Carbonates	(Soriano-Disla et al., 2014)
2130–1700	Metal–CO	(Wang et al., 2017)
2000	Quartz	(Stenberg et al., 2010)
2000–1700	Quartz and kaolinite	(Stenberg et al., 2010)
1780–1900	Metal–CO–Metal	(Wang et al., 2017)
1870	Quartz	(Stenberg et al., 2010)
1789	Quartz	(Stenberg et al., 2010)
1750–1449	Organic matter	(Janik et al., 1998)
1750–1449	Carboxylic-protein-aromatic	(Janik et al., 1998)
1725	Carboxylic acids	(Rossel and Behrens, 2010)
1720	Carboxylic acid	(Soriano-Disla et al., 2014)
1720	Carboxylic acid	(Niazi et al., 2015)
1670	Protein amide	(Soriano-Disla et al., 2014)
1666	Carboxylic acid	(Niazi et al., 2015)
1645	Water	(Rossel and Behrens, 2010)
1640	Amides	(Rossel and Behrens, 2010)

Wavenumbers (cm^{-1})	Soil constituents	References
1639	C=O	(Stenberg et al., 2010)
1633	Carboxylates	(Janik et al., 1998)
1630	Water associated organic matter	(Soriano-Disla et al., 2014)
1610	Amine	(Rossel and Behrens, 2010)
1600	Carboxylate anion	(Soriano-Disla et al., 2014)
1600–1570	Aromatic group	(Soriano-Disla et al., 2014)
1546	–NH organic acids	(Wang et al., 2017)
1530	NH+C=N, humic acid	(Stenberg et al., 2010)
1530	Protein amide	(Soriano-Disla et al., 2014)
1465	Aliphatic	(Rossel and Behrens, 2010)
1464	Aliphatic -CH stretch	(Soriano-Disla et al., 2014)
1450	Carbonate bending vibration	(Soriano-Disla et al., 2014)
1449	Alkyl	(Janik et al., 1998)
1445–1350	Methyl	(Rossel and Behrens, 2010)
1430	Carbonates	(Soriano-Disla et al., 2014)
1415	Carbonate	(Rossel and Behrens, 2010)
1413	Carboxylates	(Janik et al., 1998)
1400	Carboxylate anion	(Soriano-Disla et al., 2014)
1370	NO_3^{2-}	(Wang et al., 2017)
1350	Stretch bend of aliphatic groups	(Soriano-Disla et al., 2014)
1340–1065	C–O–H stretches	(Zimmermann et al., 2007)
1332	NO_3^{2-}	(Wang et al., 2017)
1275	Phenolics	(Rossel and Behrens, 2010)
1275	C-OH stretching in phenol	(Soriano-Disla et al., 2014)
1170	Polysaccharides	(Rossel and Behrens, 2010)
1100–1000	Quartz	(Soriano-Disla et al., 2014)
1080	Quartz	(Stenberg et al., 2010)
1080	$Si–O_2$ stretching	(Soriano-Disla et al., 2014)
1050	Carbohydrate	(Rossel and Behrens, 2010)
1050	Organic matter	(Stenberg et al., 2010)
1050	C–O stretching	(Soriano-Disla et al., 2014)
1022	Goethite or quartz	(Niazi et al., 2015)
1020	Al–OH	(Stenberg et al., 2010)
1019	Kaolinite Al–Si	(Soriano-Disla et al., 2014)
919	Al–OH	(Stenberg et al., 2010)
919	Al–OH stretching	(Soriano-Disla et al., 2014)
915	Al–OH, kaolinite	(Rossel and Behrens, 2010)
900	FeOOH	(Soriano-Disla et al., 2014)
885	Al Fe–OH, smectite	(Rossel and Behrens, 2010)
862	Al oxides	(Janik et al., 1998)
850	Al oxides	(Janik et al., 1998)
808–838	Ferrihydrite	(Janik et al., 1998)
830–730	Aromatic C-H bends	(Zimmermann et al., 2007)
824	Fe oxides	(Janik et al., 1998)
814	NO_3^{2-}	(Wang et al., 2017)
814	Goethite	(Zimmermann et al., 2007)
800	Quartz	(Stenberg et al., 2010)
800	FeOOH	(Soriano-Disla et al., 2014)
800	$Si-O_2$ stretching	(Soriano-Disla et al., 2014)
798,814	Goethite	(Niazi et al., 2015)
789	Fe oxides	(Janik et al., 1998)
764	Hematite	(Niazi et al., 2015)
706	Carbonate	(Niazi et al., 2015)
700	Quartz	(Stenberg et al., 2010)
700–600	Iron oxides	(Soriano-Disla et al., 2014)
700	$Si-O_2$ stretching	(Soriano-Disla et al., 2014)
698	SO_3^{2-}	(Wang et al., 2017)
552	SO_4^{2-}	(Wang et al., 2017)
522	Gibbsite	(Niazi et al., 2015)

Wavenumbers (cm^{-1})	Soil constituents	References
428	Si–O or Fe–O	(Niazi et al., 2015)
413	Fe–O	(Niazi et al., 2015)

Appendix 3.1 Distribution of soil samples as per NZSC soil orders (Hewitt et al., 2021) and land uses. Equivalence between NZSC soil orders and World Reference Base (WRB) soil orders (IUSS Working Group WRB, 2015) is indicative.

NZSC soil orders	WRB soil orders	Number of soil samples in each land use		
		Arable land	Non-arable land	Total
Brown	Cambisols	108	224	332
Gley	Gleysols	15	1	16
Melanic	Phaeozems	4	7	11
Organic	Histosols	2	1	3
Pallic	Lixisols	96	53	149
Podzols	Podzols	7	10	17
Raw	Regosols	6	–	6
Recent	Fluvisols	34	14	48
Semiarid	Durisols	17	14	31
Ultic	Alisols	4	–	4
Urban areas	–	–	2	2
Water bodies	–	–	3	3
Total		293	329	622

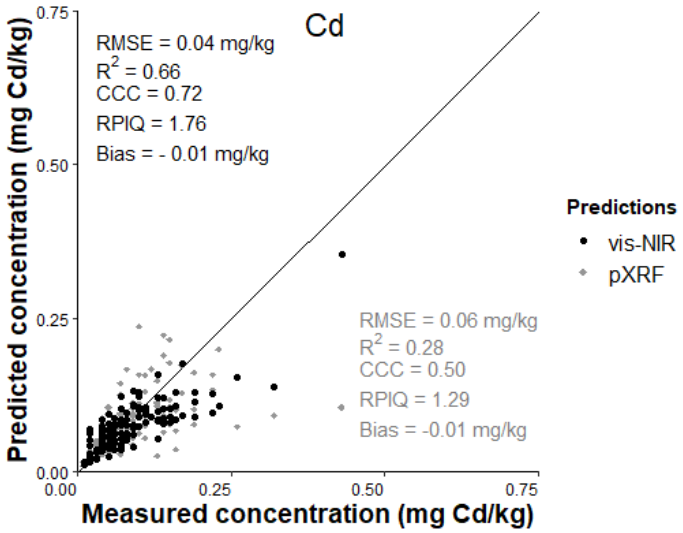
Appendix 3.2 Validation results of regression models for the testing set predicting soil TE (As, Cd, Cr, Cu, Hg, Ni, Pb, and Zn) concentrations using pXRF, vis-NIR, and MIR spectral data coupled with data fusion, model fusion, and model averaging methods.

Chemometric methods	Input spectra	As					Cd					Cr					Cu				
		RMSE	R^2	CCC	RPIQ	Bias	RMSE	R^2	CCC	RPIQ	Bias	RMSE	R^2	CCC	RPIQ	Bias	RMSE	R^2	CCC	RPIQ	Bias
PLS																					
Single sensor	pXRF	1.31	0.69	0.81	2.20	-0.12	0.06	0.18	0.42	1.12	-0.01	2.44	0.71	0.81	2.07	-0.21	2.57	0.80	0.89	2.89	0.21
	vis-NIR	1.60	0.54	0.71	1.79	-0.25	0.05	0.56	0.62	1.52	-0.02	2.67	0.64	0.79	1.89	0.03	3.43	0.62	0.77	2.17	-0.34
	MIR	1.88	0.38	0.60	1.53	-0.06	0.06	0.24	0.41	1.19	-0.02	3.19	0.48	0.65	1.58	-0.10	4.68	0.34	0.56	1.59	-0.74
Combination of sensors																					
Data fusion	pXRF + vis-NIR	1.30	0.69	0.83	2.21	0.13	0.06	0.25	0.46	1.18	-0.02	2.06	0.79	0.88	2.46	-0.17	2.36	0.82	0.91	3.15	-0.24
	pXRF + MIR	1.66	0.51	0.71	1.73	0.04	0.06	0.23	0.47	1.21	-0.01	2.67	0.65	0.80	1.89	-0.12	3.04	0.75	0.86	2.45	0.23
	vis-NIR + MIR	1.66	0.51	0.70	1.73	-0.17	0.06	0.31	0.47	1.26	-0.02	3.48	0.48	0.70	1.45	0.06	4.41	0.51	0.71	1.68	0.26
	pXRF + vis-NIR + MIR	1.49	0.62	0.78	1.93	0.12	0.05	0.35	0.53	1.36	-0.01	2.43	0.71	0.84	2.08	-0.09	3.12	0.75	0.86	2.38	0.25
Model averaging																					
GRA	pXRF, vis-NIR	1.24	0.72	0.84	2.32	<0.001	0.04	0.60	0.75	1.79	<0.001	1.94	0.81	0.89	2.60	<0.001	2.27	0.83	0.91	3.28	<0.001
	pXRF, MIR	1.28	0.70	0.82	2.24	<0.001	0.05	0.32	0.48	1.38	<0.001	2.16	0.76	0.87	2.34	<0.001	2.45	0.80	0.89	3.03	<0.001
	vis-NIR, MIR	1.57	0.55	0.71	1.83	<0.001	0.04	0.59	0.74	1.77	<0.001	2.53	0.68	0.81	2.00	<0.001	3.33	0.63	0.78	2.23	<0.001
	pXRF, vis-NIR, MIR	1.24	0.72	0.84	2.32	<0.001	0.04	0.61	0.76	1.83	<0.001	1.90	0.82	0.90	2.65	<0.001	2.26	0.83	0.91	3.28	<0.001
BMA	pXRF, vis-NIR	1.24	0.72	0.83	2.32	<0.001	0.04	0.60	0.75	1.79	<0.001	1.94	0.81	0.89	2.60	<0.001	2.27	0.83	0.91	3.28	<0.001
	pXRF, MIR	1.29	0.69	0.82	2.22	<0.001	0.05	0.32	0.48	1.38	<0.001	2.16	0.76	0.87	2.34	<0.001	2.45	0.80	0.89	3.03	<0.001
	vis-NIR, MIR	1.58	0.54	0.70	1.82	<0.001	0.04	0.59	0.74	1.77	<0.001	2.53	0.68	0.80	1.99	<0.001	3.35	0.63	0.77	2.22	<0.001
	pXRF, vis-NIR, MIR	1.24	0.72	0.83	2.32	<0.001	0.04	0.61	0.75	1.82	<0.001	1.92	0.81	0.90	2.63	<0.001	2.27	0.83	0.91	3.28	<0.001
PLS-SVM																					
Single sensor	pXRF	0.83	0.87	0.93	3.47	0.05	0.06	0.28	0.50	1.29	-0.01	1.52	0.89	0.94	3.32	0.22	1.38	0.93	0.97	5.29	0.15
	vis-NIR	1.52	0.58	0.73	1.89	-0.25	0.04	0.66	0.72	1.76	-0.01	2.40	0.71	0.82	2.11	-0.25	3.10	0.69	0.82	2.39	-0.43
	MIR	1.71	0.47	0.66	1.68	0.07	0.05	0.34	0.54	1.35	-0.01	2.95	0.57	0.74	1.71	0.16	4.21	0.44	0.64	1.76	-0.27
Combination of sensors																					
Data fusion	pXRF + vis-NIR	0.65	0.93	0.96	4.39	0.16	0.05	0.47	0.62	1.50	-0.01	1.19	0.93	0.96	4.23	-0.05	**1.09**	**0.96**	**0.98**	**6.81**	**<0.001**
	pXRF + MIR	**0.57**	**0.94**	**0.97**	**5.04**	**0.11**	0.05	0.41	0.61	1.46	-0.01	**1.14**	**0.94**	**0.97**	**4.45**	**0.09**	1.16	0.97	0.98	6.38	0.25
	vis-NIR+MIR	1.65	0.50	0.69	1.74	-0.09	0.05	0.51	0.64	1.56	-0.01	2.77	0.64	0.80	1.82	-0.13	2.74	0.76	0.87	2.71	-0.28
	pXRF + vis-NIR + MIR	1.57	0.57	0.70	1.83	-0.35	0.05	0.55	0.64	1.59	-0.01	2.54	0.68	0.80	1.99	0.27	2.81	0.75	0.87	2.64	-0.17
Model averaging																					
GRA	pXRF, vis-NIR	0.81	0.88	0.94	3.54	<0.001	0.03	0.70	0.82	2.08	<0.001	1.30	0.91	0.96	3.89	<0.001	1.48	0.93	0.96	5.03	<0.001
	pXRF, MIR	0.82	0.88	0.93	3.51	<0.001	0.05	0.44	0.61	1.52	<0.001	1.38	0.90	0.95	3.67	<0.001	1.51	0.93	0.96	4.93	<0.001
	vis-NIR, MIR	1.49	0.59	0.74	1.93	<0.001	0.03	0.70	0.83	2.09	<0.001	2.29	0.73	0.85	2.20	<0.001	3.01	0.70	0.83	2.47	<0.001
	pXRF, vis-NIR, MIR	0.81	0.88	0.94	3.54	<0.001	**0.03**	**0.73**	**0.84**	**2.17**	**<0.001**	1.28	0.92	0.96	3.94	<0.001	1.48	0.93	0.96	5.03	<0.001
BMA	pXRF, vis-NIR	0.82	0.88	0.93	3.50	<0.001	0.03	0.70	0.82	2.08	<0.001	1.30	0.91	0.96	3.89	<0.001	1.49	0.93	0.96	4.97	<0.001
	pXRF, MIR	0.82	0.88	0.93	3.49	<0.001	0.05	0.44	0.61	1.52	<0.001	1.38	0.90	0.95	3.67	<0.001	1.51	0.93	0.96	4.92	<0.001
	vis-NIR, MIR	1.50	0.59	0.74	1.92	<0.001	0.03	0.70	0.82	2.09	<0.001	2.30	0.73	0.84	2.20	<0.001	3.03	0.70	0.82	2.45	<0.001
	pXRF, vis-NIR, MIR	0.82	0.88	0.93	3.51	<0.001	0.03	0.73	0.84	2.17	<0.001	1.29	0.92	0.96	3.90	<0.001	1.49	0.93	0.96	4.97	<0.001

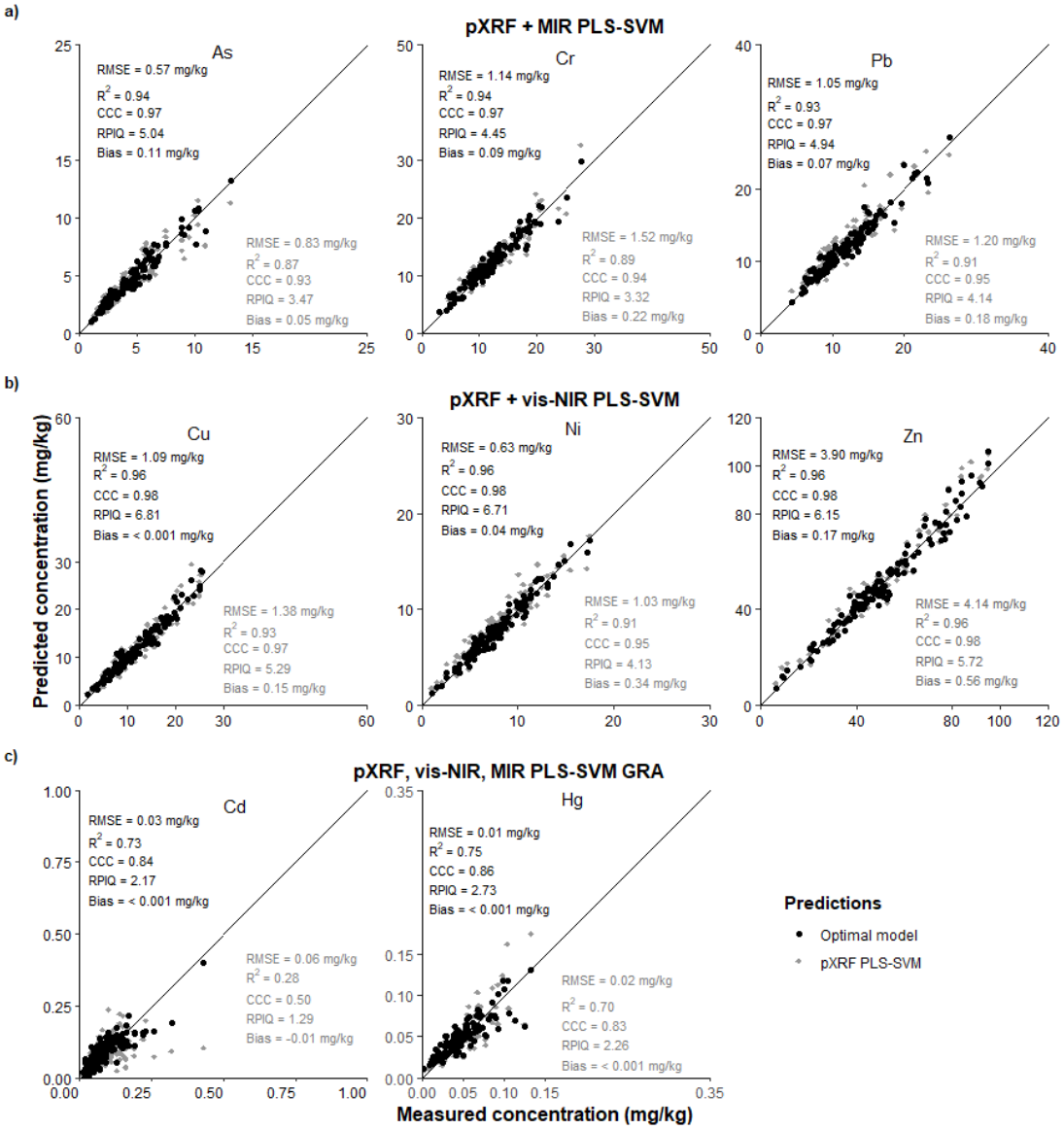
Appendix 3.2 (continued)

Chemometric methods	Input spectra	Hg					Ni					Pb					Zn				
		RMSE	R^2	CCC	RPIQ	Bias	RMSE	R^2	CCC	RPIQ	Bias	RMSE	R^2	CCC	RPIQ	Bias	RMSE	R^2	CCC	RPIQ	Bias
PLS																					
Single sensor	pXRF	0.02	0.49	0.69	1.74	<0.001	1.47	0.79	0.88	2.90	0.28	2.14	0.75	0.84	2.42	0.66	6.81	0.90	0.94	3.52	0.20
	vis-NIR	0.02	0.61	0.73	2.11	<0.001	1.76	0.70	0.83	2.41	0.13	3.42	0.30	0.50	1.51	0.12	15.92	0.42	0.59	1.51	-4.82
	MIR	0.02	0.45	0.62	1.78	-0.01	2.56	0.39	0.61	1.66	-0.10	3.45	0.31	0.51	1.50	-0.35	20.68	0.22	0.46	1.16	-1.45
Combination of sensors																					
Data fusion	pXRF + vis-NIR	0.02	0.55	0.69	1.94	-0.01	1.23	0.85	0.92	3.45	0.14	1.95	0.77	0.87	2.64	0.22	6.18	0.91	0.95	3.88	-1.16
	pXRF + MIR	0.02	0.54	0.71	1.95	<0.001	1.76	0.71	0.84	2.41	0.11	2.05	0.75	0.86	2.52	0.14	7.50	0.87	0.93	3.20	-1.29
	vis-NIR + MIR	0.02	0.61	0.75	2.12	<0.001	2.58	0.48	0.69	1.64	0.05	3.35	0.35	0.57	1.54	-0.23	17.59	0.38	0.61	1.36	-2.08
	pXRF + vis-NIR + MIR	0.02	0.60	0.75	2.06	<0.001	1.48	0.79	0.89	2.88	0.02	2.01	0.76	0.87	2.57	0.19	7.67	0.86	0.92	3.13	-1.06
Model averaging																					
GRA	pXRF, vis-NIR	0.02	0.63	0.78	2.27	<0.001	1.22	0.85	0.92	3.48	<0.001	1.98	0.76	0.87	2.60	<0.001	5.57	0.92	0.96	4.31	<0.001
	pXRF, MIR	0.02	0.59	0.74	2.13	<0.001	1.39	0.81	0.89	3.05	<0.001	2.00	0.76	0.86	2.58	<0.001	6.42	0.90	0.95	3.74	<0.001
	vis-NIR, MIR	0.02	0.63	0.77	2.24	<0.001	1.71	0.71	0.83	2.48	<0.001	3.31	0.34	0.51	1.56	<0.001	14.73	0.45	0.62	1.63	<0.001
	pXRF, vis-NIR, MIR	0.02	0.64	0.78	2.30	<0.001	1.22	0.85	0.92	3.48	<0.001	1.98	0.76	0.87	2.61	<0.001	5.52	0.92	0.96	4.35	<0.001
BMA	pXRF, vis-NIR	0.02	0.63	0.77	2.26	<0.001	1.22	0.85	0.92	3.48	<0.001	2.00	0.76	0.86	2.59	<0.001	5.57	0.92	0.96	4.31	<0.001
	pXRF, MIR	0.02	0.59	0.74	2.13	<0.001	1.39	0.81	0.89	3.05	<0.001	2.02	0.76	0.86	2.56	<0.001	6.42	0.90	0.95	3.74	<0.001
	vis-NIR, MIR	0.02	0.62	0.76	2.22	<0.001	1.73	0.70	0.83	2.46	<0.001	3.31	0.34	0.50	1.56	<0.001	14.78	0.45	0.61	1.62	<0.001
	pXRF, vis-NIR, MIR	0.02	0.64	0.77	2.27	<0.001	1.22	0.85	0.92	3.48	<0.001	2.00	0.76	0.86	2.59	<0.001	5.56	0.92	0.96	4.32	<0.001
PLS-SVM																					
Single sensor	pXRF	0.02	0.70	0.83	2.26	<0.001	1.03	0.91	0.95	4.13	0.34	1.20	0.91	0.95	4.14	0.18	4.14	0.96	0.98	5.72	0.56
	vis-NIR	0.02	0.67	0.77	2.28	<0.001	1.57	0.76	0.87	2.71	0.10	3.33	0.34	0.53	1.55	0.08	13.57	0.56	0.72	1.77	-3.19
	MIR	0.02	0.61	0.76	2.13	<0.001	2.32	0.49	0.68	1.83	-0.19	3.28	0.38	0.59	1.58	-0.42	16.28	0.38	0.60	1.47	-2.46
Combination of sensors																					
Data fusion	pXRF + vis-NIR	0.02	0.64	0.77	2.15	-0.01	**0.63**	**0.96**	**0.98**	**6.71**	**0.04**	1.15	0.92	0.96	4.48	0.09	**3.90**	**0.96**	**0.98**	**6.15**	**0.17**
	pXRF + MIR	0.02	0.66	0.79	2.30	<0.001	1.03	0.90	0.95	4.11	0.08	**1.05**	**0.93**	**0.97**	**4.94**	**0.07**	4.30	0.95	0.98	5.59	0.02
	vis-NIR+MIR	0.01	0.71	0.80	2.42	<0.001	1.67	0.74	0.86	2.54	-0.18	3.03	0.46	0.66	1.71	-0.25	12.37	0.64	0.79	1.94	-1.91
	pXRF + vis-NIR + MIR	0.02	0.69	0.81	2.39	<0.001	1.59	0.76	0.87	2.67	-0.20	3.00	0.47	0.66	1.72	-0.29	11.62	0.67	0.81	2.06	-2.01
Model averaging																					
GRA	pXRF, vis-NIR	0.01	0.74	0.85	2.68	<0.001	0.87	0.93	0.96	4.90	<0.001	1.40	0.88	0.94	3.69	<0.001	4.13	0.96	0.98	5.81	<0.001
	pXRF, MIR	0.01	0.73	0.84	2.63	<0.001	0.95	0.91	0.95	4.49	<0.001	1.42	0.88	0.94	3.63	<0.001	4.47	0.95	0.97	5.37	<0.001
	vis-NIR, MIR	0.01	0.70	0.83	2.52	<0.001	1.53	0.77	0.87	2.78	<0.001	3.16	0.40	0.57	1.63	<0.001	12.82	0.58	0.74	1.87	<0.001
	pXRF, vis-NIR, MIR	**0.01**	**0.75**	**0.86**	**2.73**	**<0.001**	0.87	0.93	0.96	4.91	<0.001	1.40	0.88	0.94	3.69	<0.001	4.10	0.96	0.98	5.85	<0.001
BMA	pXRF, vis-NIR	0.01	0.74	0.85	2.68	<0.001	0.87	0.93	0.96	4.90	<0.001	1.41	0.88	0.94	3.68	<0.001	4.13	0.96	0.98	5.81	<0.001
	pXRF, MIR	0.01	0.73	0.84	2.63	<0.001	0.96	0.91	0.95	4.44	<0.001	1.44	0.88	0.93	3.59	<0.001	4.48	0.95	0.97	5.35	<0.001
	vis-NIR, MIR	0.01	0.70	0.82	2.52	<0.001	1.54	0.76	0.87	2.76	<0.001	3.18	0.39	0.56	1.62	<0.001	12.88	0.58	0.73	1.86	<0.001
	pXRF, vis-NIR, MIR	0.01	0.74	0.85	2.72	<0.001	0.87	0.93	0.96	4.90	<0.001	1.41	0.88	0.94	3.67	<0.001	4.13	0.96	0.98	5.82	<0.001

Performance statistics results of the optimal models are in bold letters.



Appendix 3.3 Measured versus predicted Cd concentration (mg Cd/kg) for the testing set using vis-NIR (black dots; performance statistics in black letters) and pXRF (grey diamonds; performance statistics in grey letters) spectra as input for PLS-SVM.



Appendix 3.4 Measured versus predicted concentrations of a) As, Cr, and Pb, b) Cu, Ni, and Zn, and c) Cd and Hg for the testing set based on the optimal model using a combination of proximal sensors coupled with chemometric methods (black dots; performance statistics in black letters) and pXRF spectra as input for PLS-SVM (grey diamonds; performance statistics in grey letters). A black diagonal line going through the origin is 1:1 line. Spectral data fusion of pXRF + MIR as the input for PLS-SVM performed optimally for As, Cr, and Pb. Spectral data fusion of pXRF + vis-NIR as input for PLS-SVM performed optimally for Cu, Ni, and Zn. GRA of PLS-SVM model-based outputs of pXRF, vis-NIR, and MIR data performed optimally for Cd and Hg.

Appendix 3.5 Built-in pXRF calibration measurements of soil total TE (As, Cd, Cr, Cu, Hg, Ni, Pb, and Zn) for the training and testing sets.

TE	pXRF LOD (ppm)*	Training set					Testing set				
		RMSE	R^2	CCC	RPIQ	Bias	RMSE	R^2	CCC	RPIQ	Bias
As	2	5.90	0.43	0.40	0.66	4.72	5.64	0.57	0.25	0.51	5.26
Cd#	5	–	–	–	–	–	–	–	–	–	–
Cr	15	49.46	0.12	0.08	0.14	29.35	50.40	0.16	0.05	0.10	35.82
Cu	5	10.08	0.56	0.58	1.02	6.90	8.86	0.68	0.46	0.84	7.80
Hg#	1	–	–	–	–	–	–	–	–	–	–
Ni	5	10.67	0.39	0.37	0.59	7.00	10.25	0.35	0.24	0.41	7.23
Pb	3	4.90	0.27	0.50	1.19	-1.39	2.46	0.71	0.81	2.10	-1.05
Zn	3	26.49	0.58	0.53	1.19	21.05	24.26	0.81	0.58	0.99	22.16

*LOD=Limit of detection is the lowest quantity of an element in parts per million (ppm) that can be detected in an interference-free silica blank with built-in calibration in portable X-ray fluorescence spectroscopy (pXRF) spectrometer.
#All samples contained Cd and Hg concentrations well below the limit of detection of the built-in calibration in the pXRF spectrometer.

Appendix 4.1 Distribution of agricultural soil samples used in the study as per NZSC soil orders (Hewitt et al., 2021).

NZSC soil orders	Number of samples
Allophanic	50
Pumice	14
Organic	10
Pallic	6
Brown	4
Gley	2
Recent	1
Total	87

Appendix 5.1 Statistical summary of Cd concentration in soil samples included in the regional SSL, regional SSL pastoral soils subset and local set as per the NZSC soil orders (Hewitt et al., 2021).

Parameters		Allophanic	Brown	Gley	Melanic	Organic	Pallic	Podzols	Pumice	Raw	Recent	Semiarid	Ultic	Urban areas	Water bodies
							mg Cd/kg								
Regional SSL															
Number of samples	625	–	333	16	11	3	149	17	–	6	50	31	4	2	3
Minimum	0.005	–	0.005	0.02	0.03	0.03	0.005	0.01	–	0.005	0.01	0.02	0.06	0.10	0.02
1st Quartile	0.03	–	0.02	0.07	0.11	0.05	0.04	0.05	–	0.01	0.04	0.05	0.06	–	0.04
Median	0.06	–	0.04	0.10	0.13	0.06	0.07	0.08	–	0.02	0.08	0.08	0.06	–	0.05
Mean	0.08	–	0.08	0.14	0.21	0.06	0.08	0.09	–	0.04	0.11	0.08	0.07	–	0.06
3rd Quartile	0.11	–	0.09	0.13	0.29	0.08	0.11	0.10	–	0.06	0.15	0.11	0.08	–	0.08
Maximum	1.31	–	1.31	0.61	0.56	0.10	0.35	0.24	–	0.10	0.37	0.23	0.09	0.11	0.10
Skewness	5.28	–	5.94	2.27	0.85	0.09	1.79	1.22	–	0.57	1.44	0.90	0.53	–	0.16
Regional SSL pasture subset															
Number of samples	283	–	103	16	6	2	96	8	–	–	34	13	4	–	1
Minimum	0.005	–	0.005	0.02	0.10	0.03	0.02	0.03	–	–	0.03	0.02	0.06	–	–
1st Quartile	0.06	–	0.07	0.07	0.12	0.04	0.05	0.08	–	–	0.06	0.05	0.06	–	–
Median	0.09	–	0.11	0.10	0.16	0.04	0.08	0.10	–	–	0.11	0.07	0.06	–	–
Mean	0.12	–	0.14	0.14	0.25	0.04	0.09	0.09	–	–	0.12	0.08	0.07	–	0.10
3rd Quartile	0.14	–	0.18	0.13	0.34	0.05	0.12	0.10	–	–	0.16	0.11	0.08	–	–
Maximum	1.31	–	1.31	0.61	0.56	0.06	0.32	0.11	–	–	0.37	0.14	0.09	–	–
Skewness	5.36	–	4.90	2.27	0.67	0.00	1.52	-1.14	–	–	1.17	0.12	0.53	–	–
Local set															
Number of samples	87	50	4	2	–	10	6	–	14	–	1	–	–	–	–
Minimum	0.10	0.34	0.16	0.22	–	0.33	0.11	–	0.10	–	–	–	–	–	–
1st Quartile	0.35	0.55	0.22	–	–	0.47	0.11	–	0.25	–	–	–	–	–	–
Median	0.55	0.69	0.25	–	–	0.48	0.14	–	0.36	–	–	–	–	–	–
Mean	0.58	0.75	0.23	–	–	0.54	0.14	–	0.36	–	0.27	–	–	–	–
3rd Quartile	0.78	0.85	0.25	–	–	0.59	0.17	–	0.44	–	–	–	–	–	–
Maximum	2.03	2.03	0.25	0.37	–	0.87	0.18	–	0.72	–	–	–	–	–	–
Skewness	1.51	1.72	-0.72	–	–	0.63	0.05	–	0.32	–	–	–	–	–	–

Appendix 5.2 Validation results of prediction models using proximal sensor (vis-NIR, MIR, or pXRF) data of regional SSL subsets as input for PLS and LOCAL algorithms predicting total soil Cd concentration in local samples (n=87). Regional SSL subsets including R200, R250, R300, R350, R400, R450, R500, R550 were selected by PCA of each vis-NIR, MIR, and pXRF proximal sensor data from the regional SSL and local set. The regional SSL (R625) and regional SSL pasture subset (RP283) are also included.

Datasets	RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias
	PLS					LOCAL				
vis-NIR										
R200	0.52	0.18	0.84	0.08	-0.40	0.52	0.09	0.84	0.07	-0.40
R250	0.58	0.32	0.75	0.43	0.20	0.48	0.39	0.90	0.21	-0.39
R300	0.55	0.34	0.79	0.46	0.20	0.42	0.29	1.04	0.34	-0.29
R350	0.59	0.32	0.74	0.43	0.21	0.40	0.23	1.08	0.40	-0.21
R400	0.55	0.34	0.79	0.45	0.24	0.41	0.30	1.07	0.49	-0.17
R450	0.53	0.33	0.82	0.46	0.22	0.41	0.28	1.05	0.46	-0.19
R500	0.69	0.30	0.62	0.37	0.30	0.41	0.28	1.06	0.47	-0.17
R550	0.77	0.29	0.56	0.34	0.35	**0.37**	**0.32**	**1.16**	**0.53**	**-0.12**
R625	0.78	0.31	0.56	0.33	0.43	0.44	0.16	0.99	0.37	-0.16
RP283	0.47	0.30	0.93	0.47	0.16	0.45	0.08	0.96	0.28	-0.05
MIR										
R200	0.55	0.03	0.78	0.01	-0.44	0.47	0.09	0.93	0.10	-0.33
R250	0.56	0.02	0.77	0.01	-0.44	0.50	0.13	0.86	0.08	-0.38
R300	0.57	0.00	0.75	0.00	-0.46	0.54	0.02	0.80	0.03	-0.41
R350	0.57	0.00	0.75	0.00	-0.46	0.57	0.00	0.77	0.00	-0.44
R400	0.58	0.00	0.75	0.00	-0.46	0.52	0.18	0.83	0.09	-0.41
R450	0.58	0.00	0.75	0.00	-0.46	**0.46**	**0.17**	**0.94**	**0.16**	**-0.33**
R500	0.57	0.00	0.75	0.00	-0.46	0.53	0.02	0.81	0.05	-0.40
R550	0.57	0.00	0.76	0.00	-0.45	0.54	0.04	0.81	0.07	-0.40
R625	0.58	0.00	0.75	0.00	-0.46	0.59	0.00	0.73	0.00	-0.47
RP283	**0.38**	**0.33**	**1.13**	**0.21**	**-0.24**	0.51	0.05	0.86	0.07	-0.37
pXRF										
R200	0.64	0.10	0.67	0.01	-0.55	0.54	0.07	0.80	0.07	-0.42
R250	0.59	0.09	0.73	0.04	-0.49	0.54	0.05	0.80	0.05	-0.42
R300	0.64	0.10	0.68	0.01	-0.54	0.53	0.06	0.81	0.05	-0.42
R350	0.65	0.11	0.67	0.01	-0.55	0.55	0.09	0.78	0.06	-0.45
R400	0.65	0.11	0.67	0.01	-0.55	0.58	0.07	0.74	0.04	-0.48
R450	0.65	0.11	0.67	0.01	-0.55	0.58	0.10	0.75	0.04	-0.48
R500	**0.50**	**0.04**	**0.86**	**0.14**	**-0.31**	0.59	0.06	0.74	0.03	-0.48
R550	0.64	0.11	0.67	0.01	-0.55	0.59	0.06	0.74	0.04	-0.48
R625	0.52	0.04	0.83	0.16	-0.25	0.54	0.11	0.80	0.07	-0.43
RP283	0.59	0.11	0.73	0.03	-0.49	0.54	0.02	0.81	0.03	-0.41

Optimal model outputs based on land use and each sensor data are in bold letters.

Appendix 5.3 Validation results of prediction models using proximal sensor data of selected regional SSL subsets spiked with weighed local samples as input for PLS and LOCAL algorithms predicting total soil Cd concentration in local set (n=87).

Datasets/ Proximal sensors	Spiking/ Weighing	PLS					LOCAL				
		RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias
RP283											
vis-NIR	L6	0.39	0.31	1.12	0.53	0.01	0.31	0.34	1.39	0.58	-0.04
	L6×4	0.36	0.39	1.20	0.60	-0.02	0.29	0.41	1.50	0.63	-0.06
	L6×7	0.39	0.30	1.10	0.53	-0.02	0.28	0.45	1.55	0.67	-0.03
	L6×9	0.40	0.30	1.08	0.52	-0.01	0.28	0.44	1.55	0.65	-0.05
	L12	0.32	0.36	1.34	0.60	-0.01	0.28	0.48	1.57	0.66	-0.10
	L12×4	0.33	0.33	1.31	0.57	-0.04	0.26	0.53	1.65	0.68	-0.10
	L12×7	0.31	0.38	1.38	0.61	-0.06	0.28	0.51	1.57	0.66	-0.12
	L12×9	0.31	0.40	1.42	0.62	-0.07	0.27	0.52	1.60	0.67	-0.11
	L18	0.37	0.27	1.18	0.51	-0.01	0.28	0.43	1.56	0.63	-0.07
	L18×4	0.38	0.27	1.13	0.51	-0.02	0.30	0.39	1.43	0.59	-0.11
	L18×7	0.34	0.34	1.29	0.58	-0.03	0.26	0.57	1.64	0.70	-0.12
	L18×9	0.32	0.37	1.36	0.61	-0.04	0.29	0.49	1.52	0.64	-0.13
MIR	L6	0.27	0.39	1.60	0.55	0.02	0.25	0.55	1.76	0.72	-0.05
	L6×4	0.28	0.42	1.56	0.58	0.08	0.25	0.56	1.77	0.74	-0.06
	L6×7	0.27	0.43	1.58	0.58	0.08	0.25	0.54	1.74	0.72	-0.05
	L6×9	0.27	0.43	1.60	0.59	0.07	0.24	0.56	1.79	0.73	-0.05
	L12	0.27	0.39	1.58	0.52	-0.01	0.23	0.64	1.87	0.75	-0.10
	L12×4	0.27	0.42	1.60	0.52	-0.03	**0.22**	**0.66**	**1.93**	**0.78**	**-0.08**
	L12×7	0.28	0.41	1.53	0.54	-0.09	0.25	0.60	1.76	0.74	-0.10
	L12×9	0.27	0.44	1.58	0.59	-0.09	0.23	0.63	1.86	0.77	-0.09
	L18	0.28	0.38	1.53	0.47	-0.06	0.24	0.57	1.78	0.72	-0.08
	L18×4	0.29	0.41	1.50	0.45	-0.09	0.24	0.60	1.83	0.74	-0.08
	L18×7	0.27	0.49	1.60	0.60	-0.11	0.23	0.60	1.85	0.76	-0.07
	L18×9	0.26	0.52	1.64	0.62	-0.10	0.23	0.63	1.86	0.77	-0.08
pXRF	L6	0.46	0.09	0.95	0.16	-0.30	0.38	0.21	1.15	0.45	-0.07
	L6×4	0.47	0.05	0.92	0.23	-0.03	0.36	0.22	1.19	0.46	-0.07
	L6×7	0.46	0.05	0.94	0.22	0.00	0.36	0.21	1.20	0.44	-0.10
	L6×9	0.49	0.04	0.88	0.20	-0.01	0.35	0.24	1.22	0.44	-0.13
	L12	0.22	0.46	0.87	0.65	-0.03	0.33	0.31	1.31	0.51	-0.13
	L12×4	0.38	0.12	1.14	0.29	-0.14	0.35	0.25	1.23	0.46	-0.14
	L12×7	0.38	0.12	1.14	0.31	-0.13	0.34	0.30	1.26	0.46	-0.16
	L12×9	0.38	0.11	1.13	0.31	-0.11	0.34	0.32	1.27	0.48	-0.17
	L18	0.44	0.09	0.98	0.17	-0.28	0.34	0.30	1.28	0.49	-0.14
	L18×4	0.41	0.07	1.07	0.19	-0.20	0.34	0.32	1.28	0.51	-0.15
	L18×7	0.36	0.20	1.20	0.36	-0.17	0.35	0.31	1.22	0.45	-0.19
	L18×9	0.35	0.23	1.23	0.38	-0.17	0.35	0.32	1.23	0.45	-0.19
vis-NIR PCA selected R550											
vis-NIR	L6	0.54	0.31	0.81	0.45	0.16	0.32	0.44	1.34	0.59	-0.17
	L6×4	0.38	0.30	1.14	0.53	-0.08	0.31	0.49	1.39	0.62	-0.17
	L6×7	0.33	0.35	1.29	0.57	-0.10	0.29	0.54	1.47	0.69	-0.13
	L6×9	0.39	0.31	1.10	0.53	-0.10	0.32	0.48	1.36	0.64	-0.15
	L12	0.43	0.28	1.00	0.49	0.01	**0.27**	**0.59**	**1.60**	**0.69**	**-0.15**
	L12×4	0.36	0.28	1.22	0.51	-0.10	0.29	0.51	1.50	0.64	-0.15
	L12×7	0.37	0.28	1.16	0.51	-0.09	0.29	0.53	1.48	0.65	-0.16
	L12×9	0.37	0.30	1.16	0.53	-0.10	0.29	0.55	1.52	0.66	-0.16
	L18	0.41	0.31	1.05	0.53	0.05	0.31	0.44	1.42	0.59	-0.15
	L18×4	0.36	0.27	1.21	0.51	-0.06	0.27	0.57	1.59	0.68	-0.14
	L18×7	0.37	0.26	1.16	0.50	-0.06	0.28	0.56	1.54	0.67	-0.15
	L18×9	0.34	0.29	1.27	0.52	-0.08	0.30	0.46	1.43	0.62	-0.14

Datasets/ Proximal sensors	Spiking/ Weighing	RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias
		PLS					LOCAL				
MIR PCA selected R450											
MIR	L6	0.44	0.14	0.99	0.10	-0.29	0.28	0.51	1.57	0.69	-0.09
	L6×4	0.29	0.29	1.48	0.47	-0.02	0.31	0.42	1.39	0.64	-0.04
	L6×7	0.32	0.20	1.35	0.38	-0.05	0.29	0.45	1.47	0.66	-0.04
	L6×9	0.29	0.33	1.51	0.52	0.03	0.29	0.47	1.50	0.69	-0.01
	L12	0.40	0.15	1.09	0.14	-0.24	0.30	0.42	1.47	0.63	-0.06
	L12×4	0.35	0.17	1.23	0.28	-0.15	0.27	0.48	1.60	0.69	-0.05
	L12×7	0.30	0.35	1.42	0.57	-0.08	0.24	0.59	1.80	0.76	-0.06
	L12×9	0.32	0.26	1.37	0.41	-0.11	0.26	0.52	1.66	0.71	-0.07
	L18	0.39	0.16	1.10	0.15	-0.23	0.28	0.46	1.57	0.67	-0.06
	L18×4	0.34	0.28	1.26	0.34	-0.17	0.26	0.55	1.69	0.72	-0.07
	L18×7	0.29	0.43	1.49	0.62	-0.10	**0.24**	**0.62**	**1.83**	**0.77**	**-0.07**
	L18×9	0.28	0.48	1.57	0.66	-0.09	0.24	0.61	1.79	0.75	-0.08
pXRF PCA selected R500											
pXRF	L6	0.48	0.09	0.89	0.14	-0.35	0.35	0.24	1.25	0.49	0.00
	L6×4	0.49	0.05	0.88	0.23	0.01	0.38	0.18	1.14	0.42	-0.03
	L6×7	0.50	0.04	0.86	0.19	0.02	0.40	0.15	1.09	0.39	-0.04
	L6×9	0.52	0.08	0.84	0.26	0.06	0.40	0.13	1.08	0.35	-0.08
	L12	0.45	0.09	0.97	0.21	-0.26	**0.32**	**0.31**	**1.35**	**0.52**	**-0.11**
	L12×4	0.46	0.08	0.94	0.25	-0.17	0.33	0.28	1.31	0.50	-0.09
	L12×7	0.38	0.13	1.13	0.32	-0.13	0.34	0.24	1.26	0.46	-0.09
	L12×9	0.43	0.10	1.00	0.28	-0.15	0.34	0.24	1.27	0.45	-0.10
	L18	0.44	0.08	0.99	0.21	-0.23	0.35	0.30	1.24	0.49	-0.16
	L18×4	0.48	0.07	0.91	0.23	-0.18	0.35	0.30	1.23	0.49	-0.15
	L18×7	0.36	0.21	1.20	0.36	-0.18	0.35	0.34	1.22	0.50	-0.18
	L18×9	0.37	0.20	1.17	0.40	-0.15	0.36	0.31	1.20	0.49	-0.17

Optimal model outputs based on land use and each sensor data are in bold letters.

Appendix 6.1 Number of soil and plant aboveground biomass and root samples collected and analysed for Cd concentration in two glasshouse experiments.

Sample matrices	Experiments	Total number of samples collected	Number of samples analysed for total Cd
Soil	Experiment I	392	336
	Experiment II	336	112
Aboveground biomass	Experiment I	84	28
	Experiment II	82	82
Root biomass	Experiment I	28	0
	Experiment II	28	28

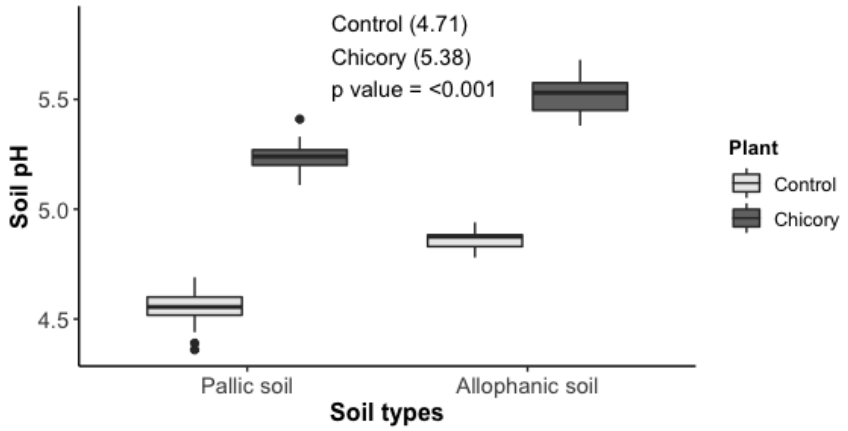
Appendix 6.2 Measured Cd concentration in soil, chicory aboveground biomass and root samples in glasshouse experiments I and II as per amended increasing soil Cd concentrations.

Added Cd concentrations (mg Cd/kg soil)	Experiment I					Experiment II				
	N	Min.	Mean	Max.	Skew.	N	Min.	Mean	Max.	Skew.
Soil (mg Cd/kg soil)										
0	48	0.17	0.21	0.27	0.37	20	0.41	0.47	0.58	0.78
0.5	48	0.52	0.61	0.71	-0.25	16	0.84	0.91	1.05	0.89
0.75	48	0.61	0.88	1.12	0.29	16	0.96	1.12	1.28	0.14
1	48	0.61	1.08	1.49	-0.44	16	1.12	1.27	1.45	0.09
2	48	1.74	2.06	2.62	1.14	16	1.89	2.10	2.69	1.74
3.5	48	2.92	3.31	3.77	0.35	16	2.85	3.40	3.87	-0.20
5	48	4.69	5.03	5.45	0.03	12	4.35	4.63	4.81	-0.33
Aboveground biomass (mg Cd/kg DM)										
0	4	3.47	3.93	4.78	0.64	12	2.40	3.16	3.97	-0.02
0.5	4	10.66	12.74	14.22	-0.37	12	5.46	10.2	26.17	1.66
0.75	4	10.05	15.81	23.33	0.35	12	6.53	12.72	21.3	0.46
1	4	19.17	25.17	33.22	0.35	12	8.52	12.42	18.57	0.26
2	4	34.27	42.34	50.02	-0.07	12	12.25	23.89	40.95	0.51
3.5	4	30.65	39.84	51.03	0.26	12	19.33	32.88	51.50	0.62
5	4	29.83	58.56	74.48	-0.59	10	25.98	40.18	56.77	0.05
***Root biomass (mg Cd/kg DM)**										
0	–	–	–	–	–	4	0.86	1.27	1.66	-0.03
0.5	–	–	–	–	–	4	2.36	2.69	3.00	-0.04
0.75	–	–	–	–	–	4	2.57	2.99	3.73	0.52
1	–	–	–	–	–	4	3.03	4.02	4.97	-0.05
2	–	–	–	–	–	4	6.07	6.68	7.66	0.41
3.5	–	–	–	–	–	4	9.04	9.39	10.15	0.70
5	–	–	–	–	–	4	16.17	19.79	25.79	0.50

*Root biomass Cd for experiment I are not available.

Appendix 6.3 Measured temporal plant aboveground biomass and root Cd concentration in the glasshouse experiments I and II amended with increasing soil Cd concentrations.

Added soil Cd concentrations (mg Cd/kg)	Experiment I	Experiment II			Root
	First	First	Second	Third	
	Aboveground biomass harvest				
	Cd concentration (mg Cd/kg)				
0	3.93[a]	2.91[a]	3.05[a]	3.52[a]	1.27[a]
0.5	12.74[ab]	7.38[ab]	8.41[ab]	14.81[b]	2.69[b]
0.75	15.81[bc]	10.60[b]	14.30[b]	13.26[b]	2.99[b]
1	25.17[c]	10.21[b]	12.58[b]	14.47[b]	4.02[c]
2	42.34[d]	16.03[c]	32.01[c]	23.63[c]	6.68[d]
3.5	39.84[d]	24.73[d]	41.71[d]	32.20[c]	9.39[e]
5	58.56[e]	28.29[d]	50.98[e]	41.00[d]	16.79[f]
Average	28.34	14.31	23.29	20.41	6.69
Added Cd concentrations effect	<0.001	<0.001	< 0.001	<0.001	<0.001
Least significant difference	11.80	5.04	9.18	8.71	0.87



Appendix 6.4 Soil pH measured for final soil samples from control (grey colour) and chicory (black) pots containing Pallic (experiment I) and Allophanic (experiment II) soils.

Appendix 6.5 Cross-validation results of PLS models predicting Cd concentration in the Pallic soil and chicory aboveground biomass samples from experiment I using laboratory vis-NIR, MIR, and pXRF spectra.

Sampling days	N	Matrices	Minimum (mg Cd/kg)	Mean (mg Cd/kg)	vis-NIR					MIR					pXRF				
					RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias
Day 0	56	Soil	0.18	1.92	0.47	0.92	5.16	0.96	-0.01	0.72	0.82	3.34	0.88	-0.16	0.42	0.93	5.80	0.97	-0.03
Day 45	56	Soil	0.19	1.90	0.35	0.95	7.34	0.97	-0.03	0.53	0.90	4.88	0.95	0.00	0.34	0.96	7.53	0.98	-0.04
Day 60	56	Soil	0.18	1.87	0.69	0.82	3.74	0.91	-0.03	0.82	0.77	3.17	0.87	-0.05	0.66	0.84	3.91	0.92	0.00
Day 73	56	Soil	0.19	1.81	0.34	0.96	7.40	0.98	-0.01	0.49	0.91	5.14	0.96	-0.02	0.35	0.96	7.22	0.98	-0.03
Day 101	56	Soil	0.18	1.63	0.36	0.95	6.99	0.97	-0.03	0.38	0.95	6.59	0.97	-0.07	0.39	0.94	6.37	0.97	-0.03
Day 127	56	Soil	0.17	1.79	0.65	0.87	3.59	0.93	0.01	0.47	0.92	5.02	0.96	-0.03	0.57	0.88	4.10	0.93	-0.07
Day 60, 73, 101, 127	224	Soil	0.17	1.88	0.42	0.93	5.94	0.97	-0.01	0.78	0.78	3.24	0.88	-0.05	0.76	0.79	3.30	0.89	-0.06
Day 0, 45, 60, 73, 101, 127	336	Soil	0.17	1.82	0.86	0.72	2.93	0.84	-0.12	1.15	0.52	2.18	0.69	-0.23	1.10	0.57	2.28	0.74	-0.17
Day 73	28	Aboveground biomass	3.47	28.34	9.60	0.81	2.77	0.90	0.17	14.89	0.47	1.79	0.58	-3.61	8.46	0.82	3.15	0.90	-0.76

Appendix 6.6 Cross-validation results of PLS models predicting Cd concentration in the Allophanic soil and chicory aboveground biomass and root samples from experiment II using laboratory vis-NIR, MIR, and pXRF spectra.

Sampling days	N	Matrices	Minimum	Mean	vis-NIR					MIR					pXRF				
			(mg Cd/kg)		RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias
Day 0	56	Soil	0.41	1.99	0.26	0.97	9.10	0.98	0.00	0.42	0.91	5.69	0.96	-0.04	0.34	0.94	7.03	0.97	-0.03
Day 21	56	Soil	0.41	1.99	0.55	0.85	4.43	0.92	-0.03	0.39	0.92	4.95	0.95	-0.06	024	0.97	9.49	0.98	-0.01
Day 0, 21	112	Soil	0.41	2.01	0.35	0.94	6.80	0.97	-0.02	0.82	0.66	2.89	0.80	-0.14	0.32	0.95	5.49	0.95	-0.06
Day 60	27	Aboveground biomass	2.41	13.79	6.87	0.47	1.88	0.67	-1.07	5.23	0.67	2.47	0.78	-1.09	3.71	0.83	3.48	0.91	-0.39
Day 97	27	Aboveground biomass	2.44	22.27	5.97	0.88	4.79	0.94	-0.77	5.27	0.91	5.43	0.95	-0.60	11.22	0.71	2.55	0.83	0.36
Day 132	28	Aboveground biomass	3.36	20.41	8.92	0.56	1.99	0.74	-1.08	8.33	0.63	2.13	0.79	-0.99	9.79	0.48	1.81	0.69	-1.25
Day 60, 97, 132	82	Aboveground biomass	2.41	18.84	7.13	0.76	2.57	0.87	-0.47	9.53	0.52	1.89	0.67	-2.40	10.49	0.46	1.71	0.64	-2.30
Day 132	28	Root	0.86	6.69	2.21	0.81	2.59	0.89	-0.44	2.32	0.72	2.09	0.81	-0.36	0.63	0.98	7.75	0.99	-0.02
Day 60 and 97 (aboveground biomass), Day 132 (aboveground biomass + root)	110	Aboveground biomass + root	0.86	15.75	6.73	0.76	2.82	0.87	-0.55	13.04	0.17	1.46	0.21	-4.02	11.31	0.37	1.68	0.46	-3.23

Appendix 6.7 Cross-validation results of PLS models predicting Cd concentration in experiment I and II combined soil and plant aboveground biomass samples using laboratory vis-NIR, MIR, and pXRF spectra.

Matrices	N	Minimum	Mean	vis-NIR					MIR					pXRF				
		(mg Cd/kg)		RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias	RMSE	R^2	RPIQ	CCC	Bias
Soil (All combined)	448	0.17	1.87	1.02	0.57	2.25	0.73	-0.17	0.76	0.75	3.02	0.86	-0.10	1.53	0.10	1.51	0.16	-0.51
Soil (Day 0)	112	0.18	1.95	0.35	0.95	6.74	0.97	-0.02	0.33	0.95	7.12	0.98	-0.01	0.91	0.65	2.62	0.79	-0.15
Aboveground biomass	110	2.41	21.26	9.87	0.68	2.19	0.82	-0.81	11.56	0.52	1.87	0.66	-3.19	7.30	0.82	2.96	0.90	-0.27

Appendix 6.8 Cross-validation results of PLS models predicting Cd concentration in soil and plant aboveground biomass samples from experiments I and II using *in-situ* vis-NIR spectra.

Experiment/ sampling day	N	Matrices	Minimum (mg Cd/kg)	Mean (mg Cd/kg)	RMSE	R^2	RPIQ	CCC	Bias
Experiment I									
Day 0	56	Soil	0.18	1.92	1.39	0.35	1.74	0.45	-0.46
Day 45	56	Soil	0.19	1.90	1.53	0.18	1.68	0.32	-0.50
Day 60	56	Soil	0.18	1.87	0.67	0.83	3.88	0.91	-0.07
Day 73	56	Soil	0.19	1.81	0.69	0.82	3.24	0.91	-0.06
Day 101	56	Soil	0.18	1.63	1.65	0.11	1.54	0.14	-0.60
Day 127	56	Soil	0.17	1.79	1.57	0.17	1.32	0.16	-0.56
Day 60, 73, 101, 127	224	Soil	0.17	1.78	1.67	0.06	1.49	0.08	-0.61
Day 0, 45, 60, 73, 101, 127	336	Soil	0.17	1.82	1.72	0.02	1.46	0.04	-0.64
Day 73	28	Aboveground biomass	3.47	28.34	15.79	0.43	1.69	0.54	-5.26
Experiment II									
Day 0	56	Soil	0.41	1.99	0.18	0.98	13.25	0.99	-0.01
Day 21	56	Soil	0.41	1.99	0.49	0.88	4.91	0.93	-0.06
Day 0, 21	112	Soil	0.41	1.99	0.31	0.95	7.61	0.97	-0.01
Day 60	27	Aboveground biomass	2.41	13.79	4.75	0.75	2.71	0.86	-0.18
Day 97	27	Aboveground biomass	2.44	22.27	5.15	0.92	5.55	0.96	-0.03
Day 132	28	Aboveground biomass	3.36	20.41	7.54	0.68	2.35	0.82	-0.76
Day 60, 97, 132	110	Aboveground biomass	2.41	18.84	5.31	0.86	3.45	0.93	-0.50

Appendix: Statement of contribution: doctorate with publications/manuscripts

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Gautam Shrestha
Name and title of main supervisor:	Professor Chris Anderson
In which chapter is the manuscript/published work?	Chapter 3
What percentage of the manuscript/published work was contributed by the student?	90%

Describe the contribution that the student has made to the manuscript/published work:

Gautam Shrestha did the laboratory scanning of soil samples, analysed data, prepared graphs, and wrote the original manuscript.

Please select one of the following three options:

- (●) **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:

  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.

- ( ) **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- ( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Gautam Shrestha Digitally signed by Gautam Shrestha Date: 2023.02.10 14:01:18 +13'00'	Main supervisor's signature:	Chris Anderson Digitally signed by Chris Anderson DN: cn=Chris Anderson, c=NZ, o=Massey University, ou=School of Agriculture and Environment, email=c.w.n.anderson@massey.ac.nz Location: Palmerston North Date: 2023.02.10 15:08:04 +13'00'

This form should be placed at the beginning of each relevant thesis chapter.

Appendix: Explanation of Covid-19 impacts

Please consider the factors listed below in your assessment of the work.

This statement has been prepared by the candidate's supervisor in consultation with the student and has been endorsed by the relevant Head of Academic Unit.

Student Name:	Gautam Shrestha	ID Number:	19034616
Supervisor Name:	Prof. Chris Anderson	Date:	8 Dec 2022
Thesis title:	Ampliyfing the power of proximal sensing techniques to assess the cadmium concentration in agricultural soils		

Considerations to be taken into account. Note: This statement will remain in the final copy of the thesis which will be available from the Massey University Library following the examination process.
[*Enter key considerations here for the examiners. This can include but is not limited to change of scope, scale, topic, focus; limitations in relation to data collection, access to necessary literature or archival materials, laboratories, field sites; disruptions as a result of lockdown and various alert levels, medical or health considerations etc*]

Gautam's glasshouse experiments and laboratory analysis were majorly disrupted by Covid-19 lockdowns and associated workplace and travel restrictions. During the first lockdown, the first glasshouse experiment that was initiated in February 2020 was abandoned. This experiment was re-initiated after an easing of NZ Covid-19 restrictions in June 2020, however the running of the experiment during winter led to a significantly slower growth period than would have been apparent during summer. Travel restrictions during the NZ lockdowns of 2021 limited opportunities to engage with instrument technicians for repairs to key analytical equipment that was needed for Cd analysis. The machine was inoperable over the period August-December 2021. Analysis was attempted using an alternate analytical instrument, however the Cd concentrations were below the limit of detection. A decision was made to complete Chapter 6 with a limited set of sample analyses obtained prior to machine failure.

Covid restrictions therefore impacted that the timeframe over which work was completed, and limited the analytical dataset on which the thesis could be written.

Approved by DRC 10/Feb/2021
DRC 21/02/03
Updated September 2021

Confidential for Examiners Only: [*Please enter any other considerations which are confidential for examiners only and should not be placed in the final thesis version submitted to Library following the examination process]*

Signed, confirming this is a fair reflection of the impact of Covid-19 on this research.

Student	Gautam Shrestha	Digitally signed by Gautam Shrestha Date: 2022.12.15 21:56:53 +13'00'
Supervisor	Chris Anderson	Digitally signed by Chris Anderson DN: cn=Chris Anderson, c=NZ, o=Massey University, ou=School of Agriculture and Environment, email=c.w.n.anderson@massey.ac.nz Location: Palmerston North Date: 2022.12.12 09:03:50 +13'00'
Head of Academic Unit (or nominee)		

Approved by DRC 10/Feb/2021
DRC 21/02/03
Updated September 2021