

Leveraging large soil spectral libraries for sensor-agnostic field condition predictions of several agronomically important soil properties

James P. Moloney^{*}, Brendan P. Malone, Senani Karunaratne, Uta Stockmann

CSIRO Agriculture and Food, Black Mountain, ACT 2601, Australia

ARTICLE INFO

Handling Editor: Jingyi Huang

Keywords:

Memory-based learning
RS-Local algorithm
Spectrum based learning
Vis-NIR spectroscopy
Soil monitoring
Soil spectral inference

ABSTRACT

Global pressures to improve soil organic carbon sequestration and soil health in general amongst the world's agricultural soils are creating a demand for improved practice to drive positive and sustainable changes in the natural capital of soils. Incentive programs aimed to promote this must be informed by accurate observations of the state of soils, both temporally and spatially. Soil spectral inference is a useful method for capturing the state of soils cost-effectively, but the price of standard laboratory grade visible and near-infrared (Vis-NIR) sensors can limit its application. Further, the acquisition of spectra by these laboratory grade sensors is performed primarily in air-dried and ground condition, adding a time lag to information retrieval. Recently, low-cost, portable miniaturised near-infrared (NIR) spectrometers have become available and have shown to be a viable alternative for the measurement of several agronomically important soil properties, which are also vital to the maintenance of soil health, including soil organic carbon (SOC), and cation exchange capacity (CEC). However, the implementation of new spectrometers, to new locations requires the creation of new spectral libraries, an expensive and labour-intensive process requiring large amounts of soil analytical and spectral data gathering. Thus, existing, laboratory grade Vis-NIR spectral libraries present a high-quality and high-resolution resource to leverage. This work demonstrates how existing spectral library resources can be accessed with cheaper, portable miniaturised NIR spectrometers with appropriate spectral filtering, and appropriate transformation matrices. In addition, the work shows that by correcting for the influences of spectral differences between soils scanned in field condition, and those prepared for analysis in the laboratory, greater uptake of spectral inference as a tool to evaluate the state of soils can be enabled. This work also demonstrates how large existing laboratory grade spectral libraries such as the CSIRO national Australian Vis-NIR soil spectral library can be queried and using memory-based learning or similar methods, such as RS-Local, and the most appropriate samples may be identified to be used for the prediction of soil properties. This work builds off an existing framework for the use of soil spectral inference for monitoring the state of soil, the Australian 2021 *Soil Organic Carbon Credits Methodology Determination*. Methods are demonstrated for the prediction of nine agronomically important soil properties, SOC, pH in water, pH in CaCl₂, electrical conductivity, CEC, and exchangeable Ca, K, Mg and Na. For SOC a model using only 20 local samples was produced in this work with a Lin's concordance correlation coefficient (LCCC) of 0.72, surpassing both the minimum requirement under the carbon credits methodology determination (LCCC 0.6), and a 50 sample local only model (LCCC 0.61). This example demonstrates that a significant further potential cost saving in laboratory analysis across soil monitoring projects can be achieved through selectively leveraging a large spectral library resource.

1. Introduction

The maintenance of soil health across agricultural environments is essential to the sustainability of soil function and to support food and fibre production for a growing global population (Herrick, 2000; McBratney et al., 2014). Monitoring the state of soil health requires

methodologies that can assess the complex physical, chemical, and biological components of the soil's condition cost-effectively and to enable plentiful sampling to address the spatially and temporally heterogeneous nature of soil landscapes. Complex biogeochemical cycling, land management activities and underlying landform and parent material variability are all important influences on state of soil

^{*} Corresponding author.

E-mail address: James.moloney@csiro.au (J.P. Moloney).

<https://doi.org/10.1016/j.geoderma.2023.116651>

Received 20 March 2023; Received in revised form 25 July 2023; Accepted 23 August 2023

Available online 13 September 2023

0016-7061/Crown Copyright © 2023 Published by Elsevier B.V. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

properties, and the dynamic nature of these influences can drive significant fluxes in the stocks of these properties. As the stocks of these properties change, so too does a soil's ability to deliver ecosystem services, improving or degrading soil health, and driving changes to soil natural capital. Successful monitoring of soil health thus requires low-cost and high throughput methodologies and tools that can inform on the stocks of these properties with consideration of the spatial and temporal variability inherent to them (Nocita et al., 2015).

In recent times, soil spectroscopy has emerged as a rapid soil measurement method as it provides many benefits over using traditional laboratory methods only, with fewer soil sample preparation steps, faster soil measurement turnaround times, lower costs and its ability to inform on multiple soil properties with a singular soil scan (Islam et al., 2003). Diffuse Reflectance Infrared (IR) spectroscopy is favoured for its ability to determine soil properties through unique reflections caused by bending and excitation of molecular bonds, which can be related to a range of soil chemical, physical and biological properties through chemometric models. Moreover, visible and near-infrared (Vis-NIR) spectroscopy offers the ability to measure soils directly in field condition (Ackerson et al., 2017). Laboratory grade Vis-NIR spectrometers including those produced by Malvern Panalytical (e.g. the ASD Fieldspec and ASD Labspec range), and Spectral Evolution (SM-3500), are currently used in research settings and commercial applications for soil property assessment. They are favoured for their resolution and wide range of operation across the electromagnetic spectrum (350–2500 nm) enabling inference of soil features that react to light in these wavelengths including soil colour, clay mineralogy, acidity, exchangeable cations, heavy metals, soil texture and organic carbon (Soriano-Disla et al., 2014; Stenberg et al., 2010). These benefits hold immense potential for real time inference, to provide rapid feedback on the states of soils. Rapid provisioning of this information can help support the uptake of decision supporting tools (Rose et al., 2016). This in turn can support informed management practices and enhanced agricultural sustainability (Bongiovanni and Lowenberg-Deboer, 2004). While these devices can provide rapid determinations of these soil properties, they are often priced well outside the limit of financial accessibility for smaller enterprises, independent consultants, or laboratories in developing economies. In contrast, lower cost near infrared (NIR) sensors have also demonstrated significant capability to predict soil properties (Soriano-Disla et al., 2017). Advances in manufacturing techniques, such as photolithography have allowed for the development of micro electro mechanical systems (MEMS), or single chip devices which miniaturise the interferometer componentry, allowing for lower production costs and smaller spectrometer form factors (Beć et al., 2020). Such instruments are generally miniaturised, and in contrast to bench top instruments described above, are often focused on the NIR region due to comparatively lower constraints in form factors for associated componentry like mirrors offered by higher brightness sources, and higher specific detectivity than other regions of the electromagnetic spectrum (Huck, 2021). Studies have demonstrated promising results with the use of MEMS NIR spectrometers for the prediction of various soil properties including soil carbon, soil organic carbon, cation exchange capacity, pH and soil texture, and the classification of soil nutrient levels (Ng et al., 2020; Tang et al., 2020).

While the lower price point can make miniaturised NIR devices attractive as alternatives to laboratory grade spectrometers, their successful performance (like any other instrument) is constrained by the requirement to establish a spectral library of sufficient quality and range (i.e., a collection of spectral responses and corresponding soil analytical values) to build robust chemometric models. Evidently, irrespective of instrumentation, the costs of the development of spectral libraries and corresponding soil measurement databases is not insignificant, but also largely unreported. Nevertheless, large spectral libraries currently exist for a variety of laboratory grade Vis-NIR spectrometers at national (Dematté et al., 2019; Shi et al., 2014; Viscarra Rossel and Webster, 2012), continental (Shepherd and Walsh, 2002), and global scales

(Batjes, 2014; Brown et al., 2006), consisting of thousands of collected spectra across many regions and environmental conditions.

Operationalising large spectral libraries has historically presented a persistent issue, as the entire breadth of spectra and analytical values contained within them may not be suitably informative at a paddock or regional level (Wetterlind and Stenberg, 2010). Strategic selection of appropriate spectra from amongst these libraries thus is an important step towards their use, leading to an increasing adoption of techniques to 'localise' models. Methods such as spiking (Barthès et al., 2020; Wetterlind and Stenberg, 2010), subset selection (Barthès et al., 2020) and techniques which constrain the library with similarity, such as memory based learning, including spectrum-based learning (Ramirez-Lopez et al., 2013), and the RS-Local algorithm (Lobsey et al., 2017). The latter two methods constrain or augment an existing library to better fit the conditions at a target site, through similarity in the spectral and/or analytical values of the local data to identify samples better aligned with the local spectra and target analyte relationships.

Large spectral libraries collected using laboratory grade spectrometers present an opportunity for easy adoption of newer, low cost, portable NIR spectrometers, particularly in cases where regions of the spectra overlap. These large spectral library resources could be exploited to reduce the challenges and investments required in construction of reference spectral libraries for new devices as they come to market, but their construction and development must be done with consideration of the portable device's utility for in situ observation. Miniaturised portable NIR spectrometers are uniquely placed for use in in situ soil conditions. Their small hand-held form, and lower power requirements lend themselves to real time use, but environmental factors such as soil moisture, temperature, humidity, and surface roughness add confounding effects to the spectra, diluting the signal of soil properties, and make spectra collected in situ incomparable to laboratory processed spectra. Further, laboratory spectral analysis, which represents the greater part of existing large soil spectral libraries, is generally conducted on homogenised, dried and ground soil samples, under uniform conditions, minimising the influences of these effects (Hutengs et al., 2019). Construction of new libraries with in situ spectra is possible, but these environmental factors vary in both space and time, and their influence on spectra is not constant, causing chemometric models produced on field acquired spectra to perform poorly when contrasted with laboratory condition soils (Waiser et al., 2007). To reduce the impact of these problems, correction methods have become available. External Parameter Orthogonalization (EPO) (Roger et al., 2003), for example, has gained popularity as the method of choice to rotate spectra against and remove undesired environmental influences. EPO was first used for soil spectral inference by Minasny et al. (2011) for the removal of soil moisture effects induced in laboratory prepared soil samples but has also been utilised to address other features of in situ soil condition including temperature, humidity, and soil matrix condition (Ackerson et al., 2017). EPO has also demonstrated utility in linking wet soil samples to existing spectral libraries (Viscarra Rossel et al., 2017). Integrating existing spectral libraries from laboratory grade sensors for in situ soil monitoring using cheaper miniaturised NIR spectrometers thus appears achievable using spectral transfer functions, and environmental correction through EPO.

In situations where legislators wish to improve environmental outcomes and ecosystem services delivered under production systems, economic-incentive-based instruments can become a primary tool (Pannell, 2008; van Grieken et al., 2019). Such programs provide commercial incentives to land managers to alter practices to benefit natural capital outcomes. Natural capital associated with soil can present an attractive target for such programs, especially with global commitments to sequester significant amounts of carbon to offset emissions (Minasny et al., 2017), but must be underpinned by assessment of the status of the soil natural capital at regular intervals. Soil focussed incentive programs that exist may combine soil observations with empirical or mechanistic models, like the Australian Commonwealth Government's *Carbon Credits*

(Carbon Farming Initiative—Estimation of Soil Organic Carbon Sequestration using Measurement and Models) Methodology Determination 2021 (Industry, Energy and Emissions Reduction, 2021) a component of the Emissions Reduction Fund (ERF) wherein management induced sequestered soil organic carbon can be credited with Australian Carbon Credit Units (ACCUs). The soil carbon accounting method within the ERF prescribes sampling events at a five-year interval, designed to capture the accumulation of soil carbon stocks through management practice change, distinct from seasonal and climactic impacts on carbon stocks (Paustian et al., 2019). The ERF soil carbon methodology is one of the first soil organic carbon methodologies to introduce the use of infrared spectroscopy to determine SOC concentrations (Environment and Energy, 2018). Spectroscopic methods are required to follow several appropriate sampling design constraints, including the creation of a moderately sized ‘local’ soil spectral library (with a required minimum of 50 samples) collected during the establishment of baseline SOC stocks. Despite the method acknowledging the benefits of leveraging a ‘large’ soil spectral library to improve prediction quality, the methodology lacks details about approaches for practical implementation of existing spectra for individual projects.

The aims of this study were therefore to construct and evaluate methods to leverage information from a large spectral library of laboratory grade Vis-NIR spectra, the CSIRO national Australian Vis-NIR soil spectral library, to predict a series of soil properties using a miniaturised portable NIR spectrometer on soils in their field condition for a local situation. Additionally, this study also aimed to demonstrate the advantage of access to an existing spectral library to enable a meaningful reduction in laboratory expenses when characterising soil natural capital using miniaturised portable NIR spectrometers, compared to existing spectroscopic measurement guidelines in a modern soil organic carbon credit program. To demonstrate utility in a real-world soil natural capital stock baseline assessment setting, this study therefore followed the guidelines of the Supplement to the Australian 2021 Soil Organic Carbon Credits Methodology Determination.

2. Materials and methods

2.1. Study area – CSIRO Boorowa Agricultural Research Station

Multiple methods for leveraging large spectral libraries for local predictions under differing spectral and sample conditions were

investigated at the newly established (2021) long-term cropping trial site (LTT) at the CSIRO Boorowa Agricultural Research Station (BARS) in southern NSW, Australia (Fig. 1). CSIRO BARS is a 220 ha mixed cropping farm and research station located close to the locality of Boorowa in south-eastern NSW, Australia. BARS belongs to the Boorowa River catchment and is characterised by gently undulating to undulating rises with a local relief ranging between 9 and 30 m above the average elevation of 600 m above sea level across the farm. BARS is characterised by a temperate climate with long summers and cool to cold winters. The farm experiences a slightly winter-dominated annual average rainfall of 619 mm. The geology at BARS comprises of Silurian ignimbrites and tuffs with interbedded sediments of the Douro group (Johnston et al., 2013). Soil types are characterised by yellow to light red duplex soils (with a texture-contrast of 20 % clay increase between the A and B soil horizon) on crests and slopes, which are classified as either Yellow or Red Chromosols or Kurosols as of the Australian Soil Classification (Isbell, 2021) depending on the presence of subsoil acidity. Other soil types found across the farm are Red and Yellow Dermosols and Kandosols and Yellow Sodosols are also found close to the drainage lines (Malone et al., 2022). The BARS LTT was identified as a location within the farm of strong utility for assessing baseline soil organic carbon distribution. The BARS LTT is an 11 ha trial site, located in the southwestern part of BARS, consisting of eight separated runs, currently zoned into four, paired farming “philosophies” under which different carbon accumulation rates would be expected.

2.2. Sampling design

Each prospective zone within the BARS LTT was treated as a Carbon Estimation Area (CEA) as of the Australian Commonwealth Government’s Carbon Credits Methodology Determination 2021 (Industry, Energy and Emissions Reduction, 2021), subdividing the trial by intended similar land management activities, with each CEA further divided into three strata. In this manner, the overall study area was subdivided into four paired CEA sections, and each was stratified independently using covariates associated with variation in soil organic carbon. Previous survey work at BARS provided a diverse set of background digital soil attribute data to stratify the BARS LTT in zones of similar conditions. Environmental covariates used in this study included gridded rasters (at 2.5×2.5 m resolution) of digital soil maps representing soil pH, texture

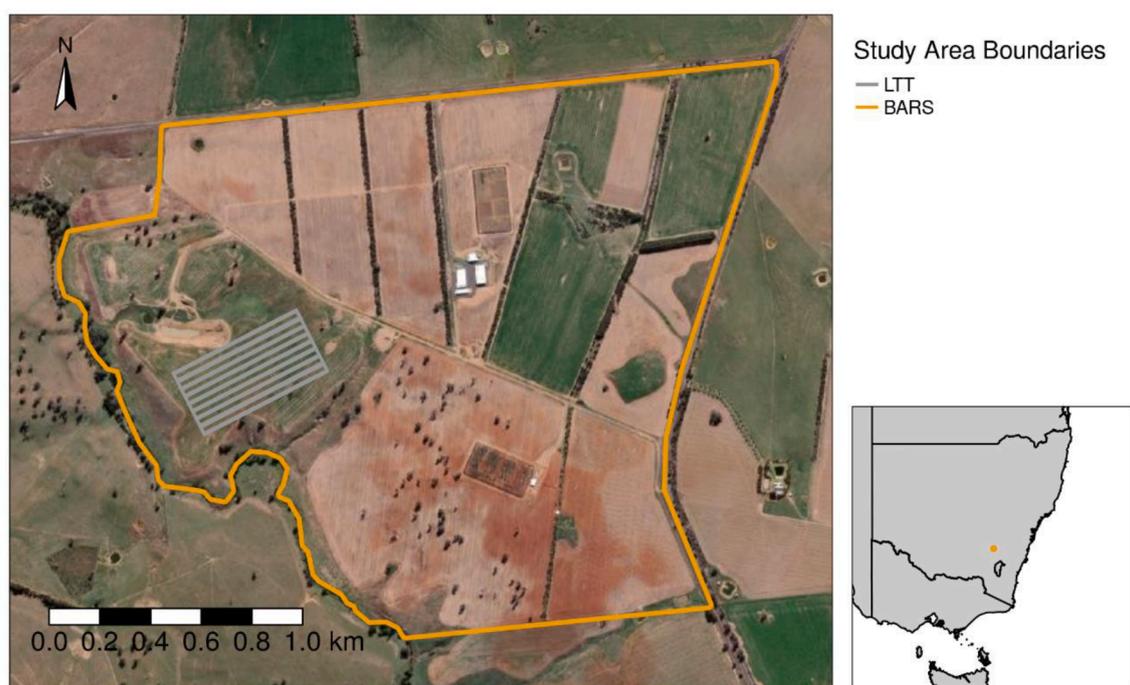


Fig. 1. A map showing the location of the CSIRO Boorowa Agricultural Research Station, and the long-term cropping trial site contained within.

and bulk density, electromagnetic conductivity and gamma radiometric surveys (Malone et al., 2022), and historic yield data, which was all supplemented with the 10th, 90th and 50th percentile normalised difference vegetation index (NDVI) values from the Copernicus Sentinel-2 MSI (European Space Agency, 2021) for the five-year period prior to sampling, retrieved from Google Earth Engine (Gorelick et al., 2017). These covariates were reprojected to a 1 m grid using a bilinear transformation, and strata were created through k-means clustering (Fig. 2). All analyses were performed using the R programming language (v4.0.2 and v4.0.5; R Core Team, 2020). For each of the four CEAs, 15 sample locations were randomly selected from across the strata within, with a minimum of three locations per stratum, and the remaining six distributed according to the size of each stratum, for a total of 60 soil sampling locations across the BARS LTT (Fig. 2). For each sampling location, two

soil cores were taken within 50 cm of each other, to a depth of 50 cm, which were sealed and refrigerated to allow for soil spectral measurements in as close to field condition as possible.

2.3. Field condition spectra acquisition – BARS LTT local spectral library

Soil cores sampled across the BARS LTT were unsealed, and the more intact core of the two was bisected to expose a large, flat surface, to increase spectrometer contact, and reduce opportunity for light leakage. Soil cores were then scanned with the hand-held Hone Lab Red NIR spectrometer in triplicate at 5 cm increments down the profile, for a total of 600 scanning locations (Fig. 3). The Hone Lab Red (Hone Carbon, Newcastle, NSW, Australia) is a miniaturised diffuse reflectance spectrometer featuring a spectral range of 1350–2500 nm, leveraging a

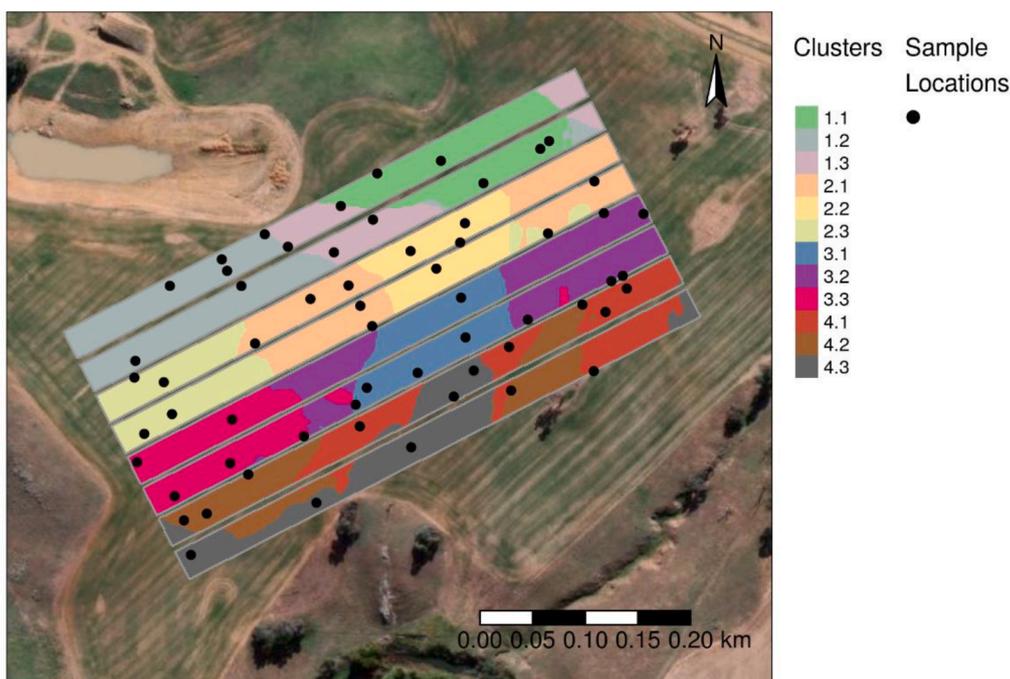


Fig. 2. A map displaying the strata generated for each of the four management zones or CEAs established within the BARS LTT, and the distribution of the sampling points within them.



Fig. 3. A diagram displaying a field condition soil core, with 10 arrows distributed in 5 cm increments to demonstrate the collection of spectral information.

NeoSpectra OEM module (Si-Ware Systems, Menlo Park, CA, United States of America) that comes at an affordable price point when compared to laboratory grade Vis-NIR spectrometers.

A total of 200 samples were selected for laboratory analysis for nine soil properties, soil organic carbon content (SOC) using dry combustion following method 6B2b (Rayment and Lyons, 2011), electrical conductivity (EC) using method 3A1, pH in water and CaCl₂ using methods 4A1 and 4B2, and cation exchange capacity (CEC) and exchangeable cations using method 15E1. Soils chosen for laboratory analysis consisted of a minimum of three sampling locations selected randomly from each core, and a further five randomly selected across each of the four CEAs.

2.4. Existing legacy spectral libraries and datasets

2.4.1. CSIRO national Australian Vis-NIR soil spectral library resource

The CSIRO national Australian Vis-NIR soil spectral library holds more than 40,000 unique spectral observations (*k*) together with information on a variety of soil analytes, with the library being collected with an laboratory grade ASD Fieldspec 3 spectrometer. For this study, all observations were extracted from amongst this library of spectra holding equivalent chemical analyses to those undertaken on the 200 samples from the BARS LTT (Table 1, Fig. 4). When possible, observations were restricted to a singular laboratory code, except in the case of SOC, which was aggregated across 5376 6A1 observations (Walkley-Black), and 602 method 6B2 observations (Dry combustion). All SOC observations were corrected to 6B2 using a conversion factor of 1.32 (Bornemisza et al., 1979). A further 11,735 EC observations in a 1:5 soil to water solution, 11,287 pH in 1:5 soil to water observations, 5744 pH in 1:5 soil to 0.01 M CaCl₂ observations, and 3218 observations of CEC and exchangeable cations (Mg, K, Na, Ca).

2.4.2. Soil water library

To address the influence of soil moisture, and surface roughness on the field condition spectra a soil water library was used, which was chosen from a set of intact soil core samples of up to 1 m of depth from the CSIRO national Australian soil archive. Using an approach outlined by Viscarra Rossel et al. (2017), 29 soil profiles were selected to represent diverse soil types and textures from regions of agricultural importance across Australia (i.e. the Harden-Young, Macquarie-Bogan Floodplain, Moree and Central West regions of NSW, as well as Emerald in QLD) (Stockmann et al., 2022). Intact soil cores were sub-sampled using a soil core smaller in diameter and of 4 cm height at three depth increments, approximately between 0 and 15 cm, 15–30 cm and 30–60 cm, to a total of 87 soil samples. Sub-sampled small soil cores were rewetted and subjected to multiple soil suctions along the soil moisture retention curve at 10 cm, 50 cm, 100 cm, 300 cm and 600 cm moisture potentials using suction plates, and at 500 kPa bar and 1500 kPa using pressure chambers at the CSIRO soil physics laboratory, ACT Australia, and following methods outlined in McKenzie et al. (2002). At each suction point, soils were scanned in triplicate with a laboratory grade

Table 1

Number of unique entries in the CSIRO national Australian Vis-NIR soil spectral library for a variety of soil properties. *Note: The EC values dataset is of low quality, with many values exceeding reasonable levels within mineral soils.

Soil Property	Chemical Method Code	Corresponding Spectra within the CSIRO Australian National Vis-NIR soil spectral library
Soil Organic Carbon	6A1	5376
Soil Organic Carbon	6B2	602
Electrical Conductivity	3A1	11735*
pH in Water	4A1	11,287
pH in CaCl ₂	4B2	5744
Cation exchange capacity and exchangeable cations: Mg, Na, K, Ca	15E1	3218

ASD Fieldspec 3 spectrometer.

2.5. Pre-processing of spectral data – CSIRO national Australian Vis-NIR soil spectral library and BARS LTT local spectral library

2.5.1. General spectral pre-processing including spectrometer harmonisation

The collected Hone Lab Red NIR spectra from all BARS LTT locations (Section 2.3), as well as the ASD Fieldspec 3 collected spectral datasets from the CSIRO national Australian Vis-NIR soil spectral library (Section 2.4.1) and the soil water library (Section 2.4.2) were pre-processed uniformly to maximise comparability of spectra. Spectra from all sources were backgrounded, ASD Fieldspec 3 spectra splice corrected at detector edges, and all spectra trimmed to the 1350–2450 nm region of the Hone Lab Red spectra (Fig. 5). Spectra were then converted to absorbance, and all spectra were treated with a 2nd order Savitzky-Golay filter with a window size of 17. The large window size of this filter was applied to address the difference in the resolution of the ASD Fieldspec 3 and Hone Lab Red spectrometers. Whilst both devices observe the spectral responses of the soil at similar wavelength spacing, with the Hone Lab Red sampling absorbance at an interval ranging from 1.2 to 4.4 and the ASD Fieldspec 3 at intervals of 1.4 and 2 nm, the actual resolution of the two spectrometers are quite different. The ASD Fieldspec 3 holds a resolution of 10 nm at the coarsest detector, whilst the Hone Lab Red observes spectra across a 16 nm window. The large window size of the Savitzky-Golay filter thus was applied to reduce any disparity in signal between the devices, by acting as a low pass filter, obscuring the high frequency signal of the ASD Fieldspec 3 spectra, and emphasising the lower frequency signal observable across both spectrometers, whilst also reducing the noise that may be imparted with finely discretised wavelength spacing. Hone Lab Red in triplicate collected spectra were visually inspected for quality control, to ensure spectral signals were consistent across replicated scans. Consistent spectra were then averaged, and these averaged spectra were utilised in all further analysis.

2.5.2. EPO generation for removal of field condition effects

To remove the influence of field condition effects, namely moisture and surface roughness, an external parameter orthogonalization matrix (EPO) was generated from the processed soil water library spectra (Section 2.4.2). For each of the 87 soil samples, the air-dry spectra were subtracted from a random moisture state selected from along the assessed points on the water retention curve, to generate a difference matrix *D*. A series of transformation matrices were generated using the EPO function in the Soilspec package in R (Wadoux et al., 2021), from difference matrix *D* and different numbers of principal components, from which the Wilkes lambda was calculated. A maxima in Wilkes lambda was identified at four principal components, and thus this value was used in the generation of the final transformation matrix.

2.6. Soil property prediction model generation

To assess the suitability of the generated EPO for moisture correction on spectra collected under a different spectrometer, chemometric models were produced, both with, and without transformation of the spectra. Cubist, a rule-based data mining algorithm was used to create calibration models for the prediction of nine agronomically important soil properties, SOC, pH in water, pH in CaCl₂, electrical conductivity, CEC, and exchangeable Ca, K, Mg and Na. Cubist is a regression rule model and is an extension of Quinlan's (1992) M5 model tree. During model building one or more rules are created, with each rule being a linear model of the predictors, with the possibility of rules overlapping. Cubist was executed using the Cubist package in R. We implemented a modelling approach employing bootstrap resampling with replacement, over 50 bootstraps or bags, and predicted values were reported as the mean predicted value across all bootstraps. This bootstrapping method

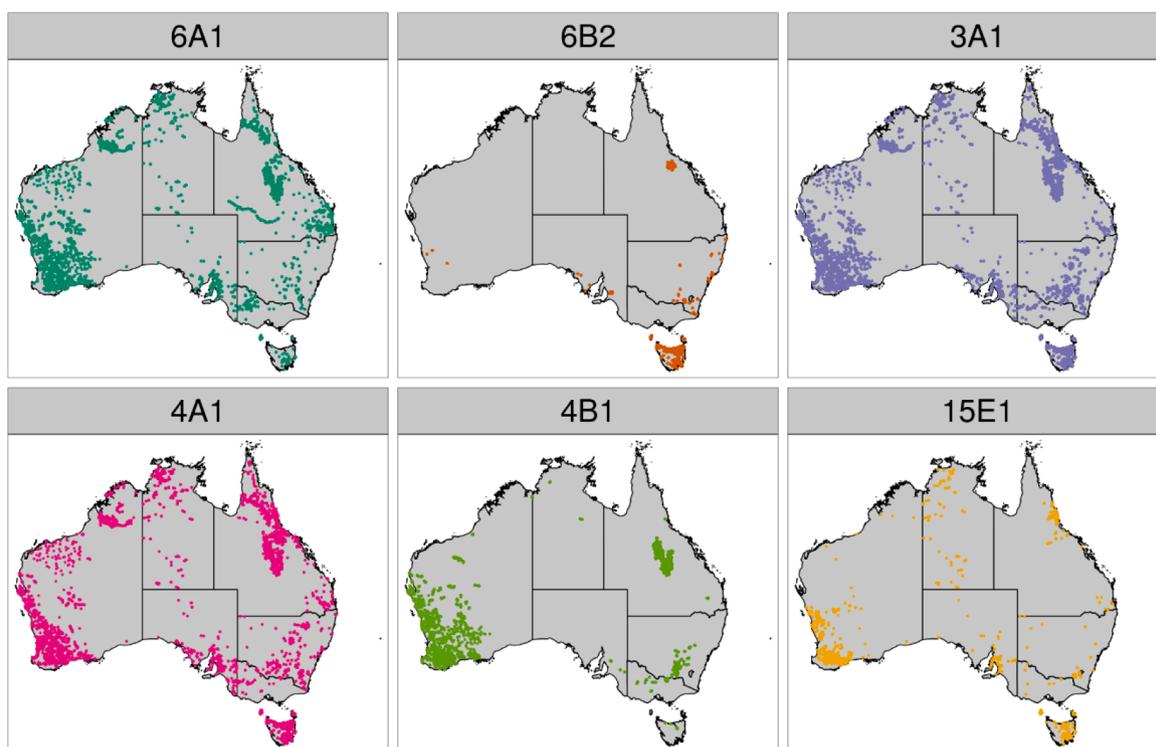


Fig. 4. Facetted maps displaying spatial distribution of entries in the CSIRO national Australian Vis-NIR soil spectral library for a variety of soil properties examined within this work. Samples located in Australian overseas territories have been omitted for clarity.

allows for quantification of the uncertainty of a prediction, through the generation of a distribution of model outcomes for each sample from which confidence intervals can be computed (Malone et al., 2022; Viscarra Rosset et al., 2015), though these uncertainties are not explored within this work. Models were trained on 50 samples selected through conditioned Latin hypercube sampling (cLHS) (Minasny and McBratney, 2006) of the spectral space, using the cLHS package in R (Roudier, 2011). Where soil property data were skewed or non-normal, as was the case for all but our pH observations, a natural log transformation were applied. A series of goodness of fit metrics were reported on the back

transformed predicted values from our bootstrap model for both the calibration dataset, and the remaining 150 unused data points as a validation set. Model evaluation metrics included the coefficient of determination (R^2), root mean square error (RMSE) and bias as well as the Lin's concordance correlation coefficient (LCCC) (Lin, 1989) and the ratio of performance to interquartile range (RPIQ) (Bellon-Maurel et al., 2010).

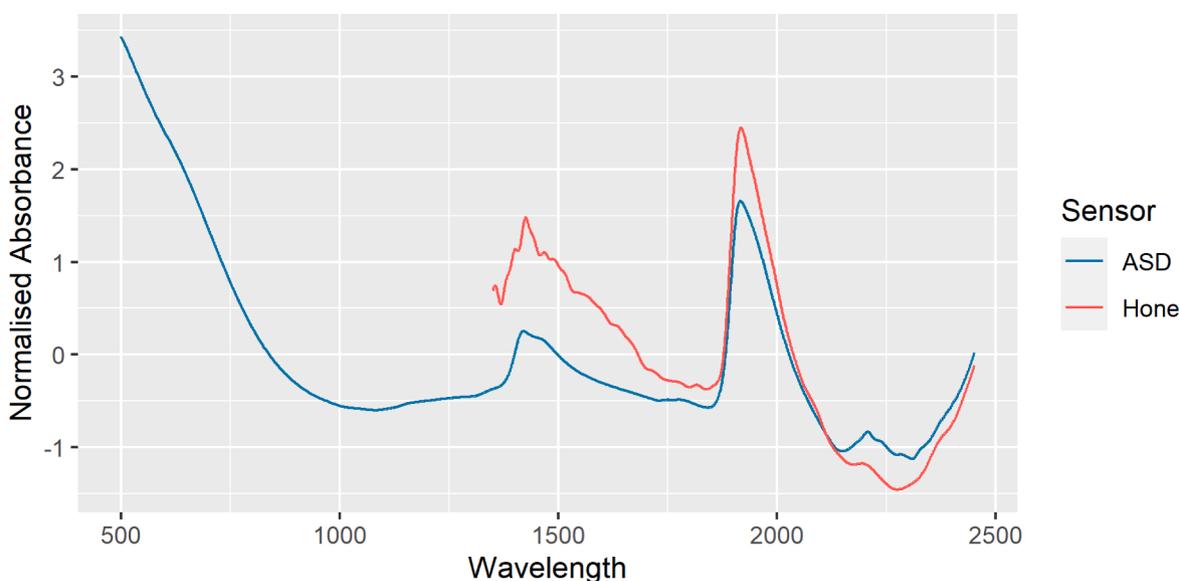


Fig. 5. A comparison of absorbance spectra treated with a standard normal variance transformation for both an ASD Fieldspec 3 and a Hone Lab Red spectrometer for an example soil specimen, displaying the difference in spectral region covered by each spectrometer between 500 and 2450 nm and the obscuration of features by the lower-resolution Hone Lab Red.

2.7. Library constraining approaches used for large spectral library selection for the local situation

To identify solutions for leveraging large spectral libraries under disparate spectral and sample conditions, two approaches which constrain a spectral library to a local observed situation were implemented, each utilising varying amounts of local data for prediction

model generation, to evaluate the utility of supplementary large spectral library samples (Fig. 6).

2.7.1. Adapted RS-Local

The first method, an adapted implementation of the RS-Local approach (Lobsey et al., 2017) is designed to select those samples from amongst a large spectral library that perform the best when

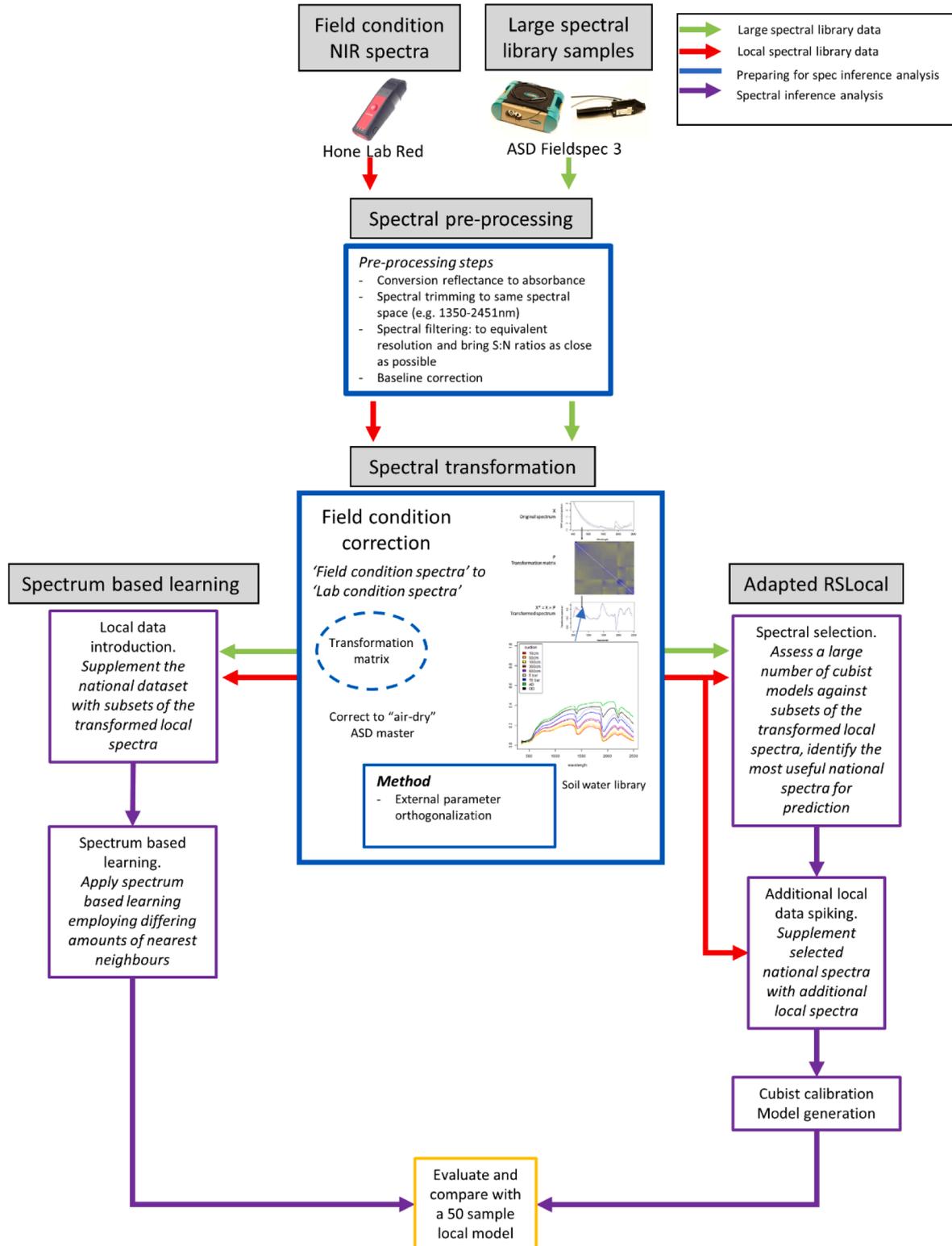


Fig. 6. Workflow diagram demonstrating the treatment of spectra and execution and evaluation of the two library constraining approaches examined.

predicting the response variable for a training set m . To that end, the method generates many spectroscopic models based on random small subsets of a spectral library and assesses the performance of those models against the training set m . For each of these random models, errors are then recorded alongside the selected samples. Following a sufficiently large number of randomly sampled models, a distribution of residuals is created, and samples from amongst the library included consistently in those models with the lowest residuals selected out from amongst the greater library. This approach localises the library to observed conditions.

From amongst the local BARS LTT soil spectral library (i.e. samples that were selected for laboratory analysis that hold a spectral response and corresponding soil analyte values), a series of site-specific comparison sets m were selected using cLHS. A total of 10 different comparison sets m were chosen, of sizes ranging from 5 to 50 samples, incremented by 5 samples each time. For each of these comparison sets m , the RS-Local methodology was adapted, wherein k samples were selected from the large CSIRO national Australian Vis-NIR soil spectral library K , and used to construct a chemometric model to predict one of the nine target soil properties, with the performance of this model evaluated against comparison set m , and RMSE recorded. This process was repeated many times (B), and the RMSE values were assigned cumulatively to each selected sample in each k subset, and once B repeats were finished the locations with the lowest cumulative RMSE values were retained. The original RS-Local method was modified slightly, substituting Cubist prediction models for partial least squares regression, and selection of the locations with the lowest median RMSE predictions, rather than the lowest cumulative RMSE. A consistent value of 300 for k was used across all nine soil properties assessed, and a B value of 5000. For each soil property and each unique comparison dataset m , a localised spectral library of the 300 samples with the lowest median RMSE, plus the initial selected samples m was created, and used to construct a bootstrapped Cubist model. Further, the addition of spiking samples was assessed, selected from amongst the remaining laboratory-analysed soils using cLHS with consideration of the region of spectral space currently represented by the comparison set m , and the RS-Local selected national spectra. Spiking samples were added to these assemblies of localised libraries and comparison sets m in groups of five, until the total number of local samples used in the construction of a model reached 75, and bootstrapped Cubist models were again

produced for each of these spiked datasets. All models were validated using hold-out validation on the remaining laboratory analysed samples, i.e. if a model was generated using an m set of 5, and 10 additional local samples were used to spike the selected national spectra, the prediction performance was reported on the remaining 185 analysed samples. This resulted in differing validation sets for as the selected spiking samples were dependant on the spectral space of both the training set m and the identified national spectra. LCCC and RMSE values amongst from each of these models were reported.

2.7.2. Spectrum-based learning

Leveraging the CSIRO national Australian Vis-NIR soil spectral library was also explored with spectrum-based learning (Ramirez-Lopez et al., 2013). Spectrum-based learning uses a multiple step process to provide unique predictions of each point within a target set (m), by first recovering the k most spectrally similar locations from a large library of spectra (Yr) and set of associated analytical values (Xr). Models are fit using these spectrally similar datapoints and associated analytical values for each point, and finally the predictions are generated for the unseen analytical values (Xu) of m . As with the adapted RS-Local approach, the addition of 10 different sets of local spectral information (BARS LTT soil spectral library) to the national set was investigated, essentially increasing Yr and Xr with supplementary local observations, ranging in size from 5 to 50 samples (XYr_s) (Fig. 7). The local samples included in XYr_s selected were identified through cLHS of the spectral space of the local scans, and the resemble package (Ramirez-Lopez et al., 2016) in the R programming language was used to iterate through a variety of different values of k , between 50 and 300 samples identified by low Mahalanobis distances, in increments of 50. PLSR models were employed to generate predictions of holdout local spectra for each XYr_s and k value, as the method available within the resemble package for the calibration of local models with the most widespread adoption, and the LCCC and RMSE values from each of these different sets was reported.

2.7.3. Assessing library constraining methodologies

To assess their potential utility against a real-world soil spectral inference framework for soil stock accounting, the performance of each library constraining approach to select appropriate samples from the large CSIRO national Australian Vis-NIR soil spectral library for the BARS LTT local situation, was contrasted against prediction model

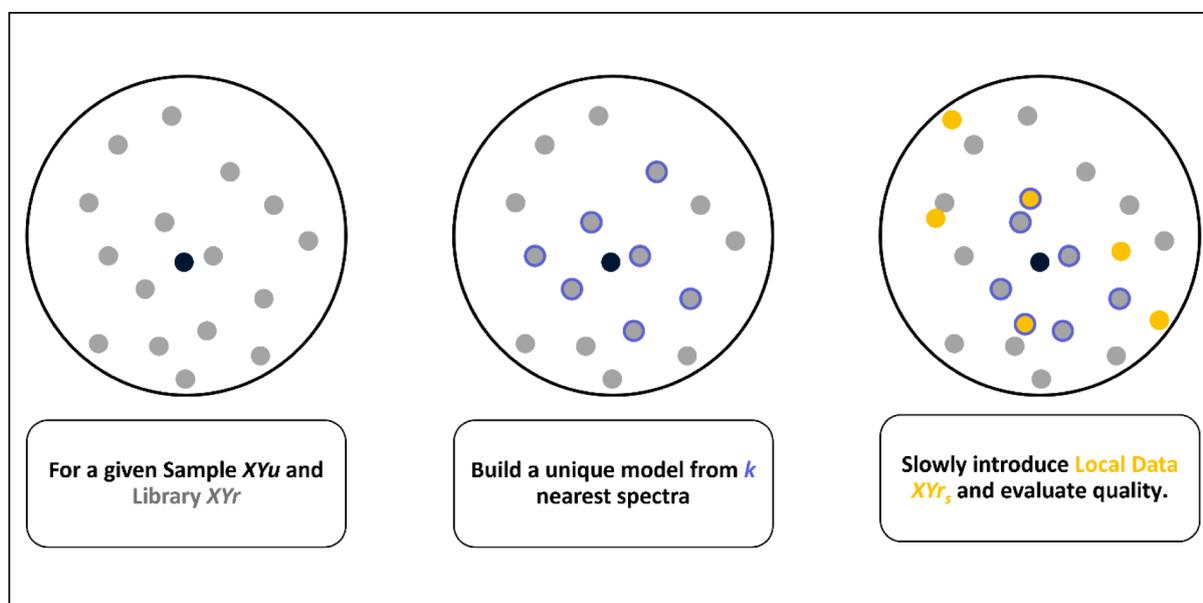


Fig. 7. Conceptual model demonstrating the process of producing predictions using spectrum-based learning, and the addition of local spiking samples, XYr_s , to investigate the use of a large spectral library for local-scale predictions.

performance emulating the requirements of the Australian Government ERF Soil Carbon Methodology 2021 spectral inference approach. The soil carbon method within the ERF leverages soil spectral inference, combined with stratified random sampling and bulk density estimation to produce an unbiased estimation of the soil carbon stocks over an area, for the purposes of identifying offset carbon emissions, and provision of carbon credits. A soil carbon project, can, as is the case in this study, consist of multiple carbon estimation areas under which different management practices or rates of carbon sequestration are expected to occur. Under the methodology, each project requires a minimum soil spectral library size of 50 new samples. Whilst the use of legacy soil spectral information is encouraged, there are no clear provisions as to the how within the 2021 methodology determination. At present, the methodology holds very few model quality requirements that relate specifically to the prediction model quality of the spectral inference, with the singular requirement that a model must have a Lin's concordance correlation coefficient of 0.6 or greater. As this is a singular dimension of model quality assessment, and because an LCCC of 0.6 is not a universally achievable metric with NIR spectroscopy across all of the chemical analytes assessed within this work (Viscarra Rossel and McBratney, 2008), in addition to the requirements as of the Australian Government ERF Soil Carbon Methodology 2021, we report the RMSE values of the predictions. Model quality was primarily assessed against the validation performance of a 50 sample local bootstrapped Cubist model on the remaining 150 samples within the BARS LTT local library that were not used in model training.

3. Results and discussion

3.1. Evaluation of BARS LTT local only soil property prediction models generated with or without EPO transformation

The 50 sample calibration model fits of the field condition spectral response to the measured soil analytical properties were generally of high quality, both with, and without transformation by EPO (Table 2). Across all properties, R^2 values and LCCCs exceeded 0.8, indicating a high degree of fit of predicted and observed values of SOC, EC, pH in water and CaCl_2 , CEC and exchangeable cations. The RPIQ however demonstrates only moderate calibration performances, with relatively large values of RMSE to the generally narrow distributions of local soil properties (Fig. 8). Further, most goodness of fit metrics improved for the models created on field condition soil samples when spectra were transformed using the EPO, demonstrating that the confounding influences of soil moisture and surface roughness upon the soil spectra can be successfully reduced. This also demonstrates that the EPO generated

from spectra collected on the higher resolution ASD Fieldspec 3 spectrometer and transformed toward the lower resolution Hone Lab Red can be used with confidence.

The improvement in prediction quality provided by the EPO, persisted when the model quality was assessed under holdout validation, except for the exchangeable Mg predictions, where the uncorrected spectra provided better predictions. There was a general reduction in the quality of fit metrics across the holdout validation predictions compared to the calibration sets. The best performing models were those of CEC, exchangeable Ca and Mg, followed by SOC, with LCCC values exceeding 0.6 in each of these properties, the minimum acceptable standard for use within a spectroscopic model under the ERF Soil Carbon Methodology 2021. The overall model fit metrics observed in validation were moderate, with the greatest performing model, EPO corrected CEC, holding an R^2 value of 0.74, and an LCCC value of 0.81 (Table 3), indicating a quality fit along a 1:1 observed to predicted line, but an RPIQ value of 0.87, showing a lower consistency of the predictions. The same holds true for the predictions of exchangeable Mg, where both EPO treated and untransformed predictions held high LCCC values of 0.79 and 0.82, respectively, but relatively large RMSE values resulted in low RPIQ values of 0.25, and 0.37. All other models held RPIQ values lower than 1, even where R^2 and LCCC values remained high. This highlights persisting limitations in providing consistent prediction of soil properties from field condition soils with lower resolution, handheld miniaturised NIR sensors when using small training sets (50 samples).

3.2. Evaluation of the soil property and spectral range present in the BARS LTT local and compared to the CSIRO national Australian Vis-NIR soil spectral library

The samples analysed from the BARS LTT had an analytical value range that was generally well represented within the CSIRO national Australian library, with all values falling well within the range of existing soil observations across the almost 40,000 combined spectra and analytical values (Fig. 8, Table 1). Most soil properties saw a large overlap between the 1st and 3rd quartiles of the local and CSIRO national Australian libraries, with exceptions generally occurring for the exchangeable cations, where the local and national values of exchangeable Mg and Na were quite different, with very low exchangeable Na observations in the local BARS LTT dataset with many samples below detection limit, and large exchangeable Mg values, with a median value of $0.56 \text{ cmol} + \text{kg}^{-1}$ compared to the national median of $0.09 \text{ cmol} + \text{kg}^{-1}$.

Spectral similarity was assessed using the principal component space of a combined dataset of both the transformed, field condition Hone Lab

Table 2

Calibration goodness of fit metrics for 50 sample local Cubist models generated on field condition scans of soils sampled from the BARS LTT, both corrected and uncorrected by external parameter orthogonalization.

Property	Units	Moisture Correction	R2	LCCC	RMSE*	Bias*	RPIQ
Carbon	%	EPO	0.90	0.91	0.10	0.00	0.81
Carbon	%	Untransformed	0.87	0.86	0.12	-0.01	0.66
EC	dSm^{-1}	EPO	0.93	0.92	0.01	0.00	1.36
EC	dSm^{-1}	Untransformed	0.90	0.90	0.01	0.00	1.22
pH water		EPO	0.90	0.84	0.30	0.01	0.93
pH Water		Untransformed	0.88	0.79	0.33	-0.01	0.82
pH CaCl_2		EPO	0.88	0.87	0.29	-0.01	1.36
pH CaCl_2		Untransformed	0.84	0.80	0.35	-0.02	1.14
CEC	$\text{cmol} + \text{kg}^{-1}$	EPO	0.96	0.95	0.82	0.05	1.65
CEC	$\text{cmol} + \text{kg}^{-1}$	Untransformed	0.95	0.94	0.89	-0.08	1.51
Exch Ca	$\text{cmol} + \text{kg}^{-1}$	EPO	0.94	0.92	0.70	-0.07	1.14
Exch Ca	$\text{cmol} + \text{kg}^{-1}$	Untransformed	0.91	0.89	0.83	-0.09	0.97
Exch K	$\text{cmol} + \text{kg}^{-1}$	EPO	0.88	0.86	0.08	0.00	1.02
Exch K	$\text{cmol} + \text{kg}^{-1}$	Untransformed	0.84	0.88	0.07	0.01	1.05
Exch Mg	$\text{cmol} + \text{kg}^{-1}$	EPO	0.97	0.95	0.35	-0.03	0.56
Exch Mg	$\text{cmol} + \text{kg}^{-1}$	Untransformed	0.97	0.93	0.41	-0.05	0.47
Exch Na	$\text{cmol} + \text{kg}^{-1}$	EPO	0.93	0.88	0.06	0.00	0.00
Exch Na	$\text{cmol} + \text{kg}^{-1}$	Untransformed	0.90	0.85	0.06	0.00	0.00

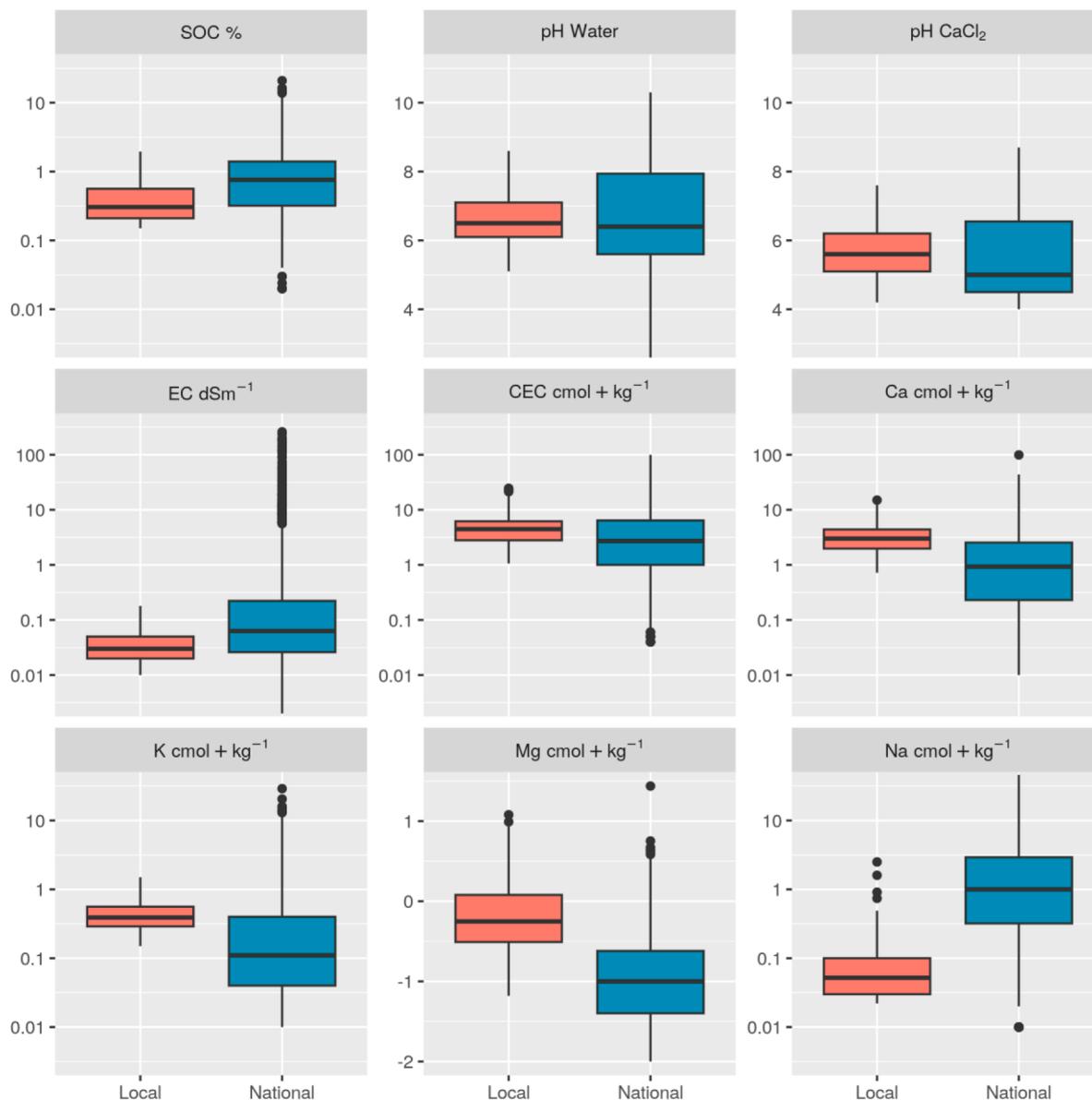


Fig. 8. Boxplots of the distribution of the nine soil properties investigated at the BARS LTT, and the distributions of the same properties within the CSIRO national Australian Vis-NIR soil spectral library. All properties except pH in water and pH in CaCl₂ are presented in Log₁₀ transformation.

Table 3

Hold-out validation goodness of fit metrics for 50 sample local Cubist spectral models generated on field condition scans of soils sampled from the BARS LTT, for SOC, EC, pH, CEC and exchangeable cations contrasting the use of an EPO transformation.

Property	Units	Moisture Correction	R2	LCCC	RMSE*	Bias*	RPIQ
Carbon	%	EPO	0.49	0.62	0.25	-0.02	0.44
Carbon	%	Untransformed	0.48	0.53	0.27	-0.06	0.41
EC	dSm ⁻¹	EPO	0.34	0.50	0.03	0.00	0.39
EC	dSm ⁻¹	Untransformed	0.24	0.40	0.03	-0.01	0.36
pH water		EPO	0.28	0.39	0.62	-0.09	0.92
pH Water		Untransformed	0.32	0.38	0.62	-0.15	0.92
pH CaCl ₂		EPO	0.30	0.47	0.66	-0.12	0.87
pH CaCl ₂		Untransformed	0.34	0.47	0.64	-0.12	0.89
CEC	cmol + kg ⁻¹	EPO	0.74	0.81	2.07	0.07	0.87
CEC	cmol + kg ⁻¹	Untransformed	0.70	0.80	2.15	-0.09	0.84
Exch Ca	cmol + kg ⁻¹	EPO	0.57	0.66	1.64	0.04	0.72
Exch Ca	cmol + kg ⁻¹	Untransformed	0.53	0.64	1.69	-0.08	0.69
Exch K	cmol + kg ⁻¹	EPO	0.35	0.41	0.21	-0.01	0.53
Exch K	cmol + kg ⁻¹	Untransformed	0.35	0.52	0.21	0.01	0.55
Exch Mg	cmol + kg ⁻¹	EPO	0.64	0.79	1.07	0.17	0.25
Exch Mg	cmol + kg ⁻¹	Untransformed	0.73	0.82	0.85	0.00	0.32
Exch Na	cmol + kg ⁻¹	EPO	0.33	0.49	0.21	0.01	0.00
Exch Na	cmol + kg ⁻¹	Untransformed	0.33	0.38	0.21	-0.01	0.00

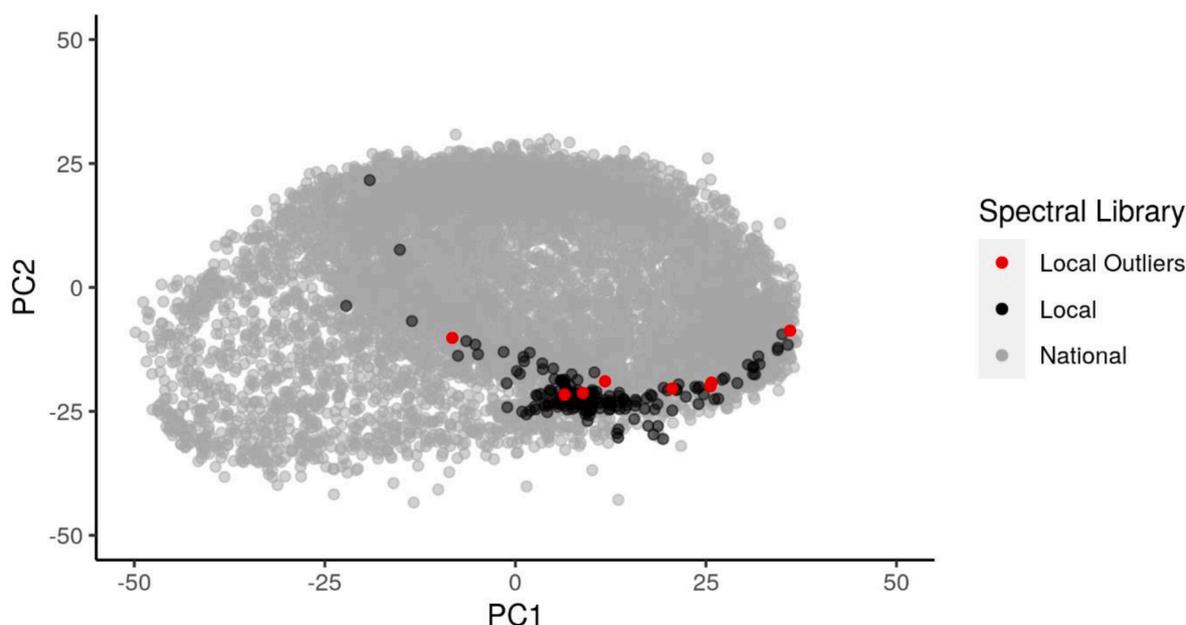


Fig. 9. Biplot of the first two principal components generated from the EPO treated spectra from both the local, field condition spectra, and laboratory condition spectra from the CSIRO national Australian Vis-NIR soil spectral library. Spectra originating from the national library are presented in grey, and local spectra in black. Outliers from amongst this combined distribution which originate from the local spectra are identified in red.

Red BARS LTT local spectra, and the CSIRO national Australian Vis-NIR soil library spectra. Using the first 8 principal components, which represented more than 95% of the variance of the combined spectral library, the Mahalanobis distances were calculated, and outliers assessed as falling above the 97.5 % quantile of the χ^2 distribution. Of those outliers, only 8 were located within the BARS LTT spectral library, indicating a decent overlap between the spectral space of each of the two EPO transformed spectral libraries (Fig. 9).

3.3. Evaluation of leveraging the CSIRO national Australian Vis-NIR soil spectral library using the adapted RS-Local approach

Predictions generated from the various localised libraries (i.e. selected national and differing amounts of local library data combined), and associated Cubist models generated through RS-Local were generally quite proficient at predicting SOC. Predictions showed a general trend of increasing quality, as the amount of available local data

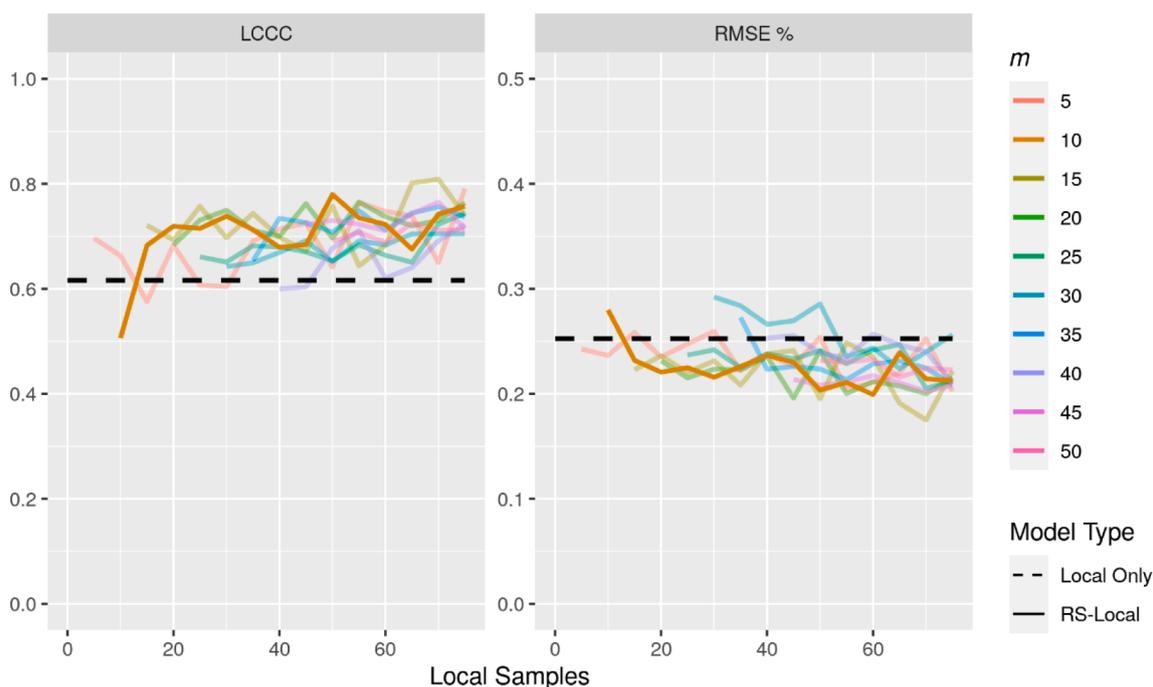


Fig. 10. LCCC and RMSE values for RS-Local selected national and local combined model predictions of soil organic carbon generated from the filtered and EPO-transformed CSIRO national Australian Vis-NIR soil spectral library using differing sizes of comparison set m , represented by differing colours and additional local data gathered in field condition. The corresponding goodness of fit metrics from a 50 sample BARS LTT local only model are presented as a dashed black line. Highlighted here is the smallest m sets which, when supplemented with additional spiking samples, consistently achieves LCCC higher, and RMSE values lower than the 50 sample local only reference model (i.e. the brown line with an m set of 10 and 5 local samples).

increased, with decreases in RMSE, and increases in LCCC. Only 6 models displayed LCCC values below the 0.62 (Fig. 10) of the BARS LTT local only EPO transformed and moisture corrected 50 sample model (Table 3), with national and local combined models sourced from RS-Local libraries trained on m values of 5, 10 and 40 local samples. RMSE values also largely fell below the values from the BARS LTT local only 50 sample, moisture corrected model, with 11 models exceeding 0.27 % (Table 3). Generally, the BARS LTT local only model was quickly surpassed, with a national and local combined model employing an m of 10 samples showing a surpassing of the BARS LTT local only model at 5 additional spiking samples, and achieving an LCCC of 0.72 with 10 additional spiking samples. This resultant LCCC showed a high fit of observed and predicted SOC values, using only 20 local samples. Of interest, some larger values of m seemed to hold higher RMSE when there were few or no additional spiking samples, seen in m values of 30 and 40 where national and local combined models constructed with up to 20 additional local samples in addition to those used in m held RMSE values greater than the 50 sample BARS LTT local only model. This lag in the reduction of RMSE may suggest that larger values for m require larger amounts of additional local data to overcome overfitting in the selection of appropriate spectra and corresponding analytical values through RS-Local.

For the models of pH in water and CaCl₂ from the RS-Local localised libraries, similar improvement in LCCC and RMSE values to the BARS LTT local only 50 sample models as seen when using smaller amounts of

local observations in the SOC predictions were not observed. For pH in CaCl₂, a similar prediction quality to the BARS LTT local only 50 sample model was approached using fewer local observations. The first RS-Local selected national and local combined model to approach the quality of the BARS LTT local only model was constructed using an m set of size 35 spiked with an additional 5 samples, for a total of 40 local observations. Here, an LCCC of 0.45 and an RMSE of 0.63 were achieved (Fig. 11), each just below those values seen in the BARS LTT local only model for pH in CaCl₂ (Table 3). Whilst this does represent a reduction in fit of observed to predicted pH values along the 1:1 line, this RS-Local selected national and local combined model holds slightly higher accuracy than the BARS LTT local only model. Unlike the pH in CaCl₂ models, no pH in water model approached the goodness of fit metrics of the 50 sample BARS LTT local only model using fewer local observations. The first pH in water RS-Local selected national and local combined model to approach the fit metrics of the 50 sample BARS LTT local only model was constructed again using an m set of size 35, but here with 20 additional spiking samples, for a total of 55 local observations. The LCCC of this model was slightly above that of the 50 sample local only model, at 0.39, and the RMSE was slightly lower at 0.6, however, as the total number of local observations exceeded those of the corresponding local model, this is somewhat expected.

Generally, similar patterns were seen in the prediction of EC (dS m⁻¹) to that seen in the prediction of pH in CaCl₂, with RS-Local selected national and local combined model quality only approaching

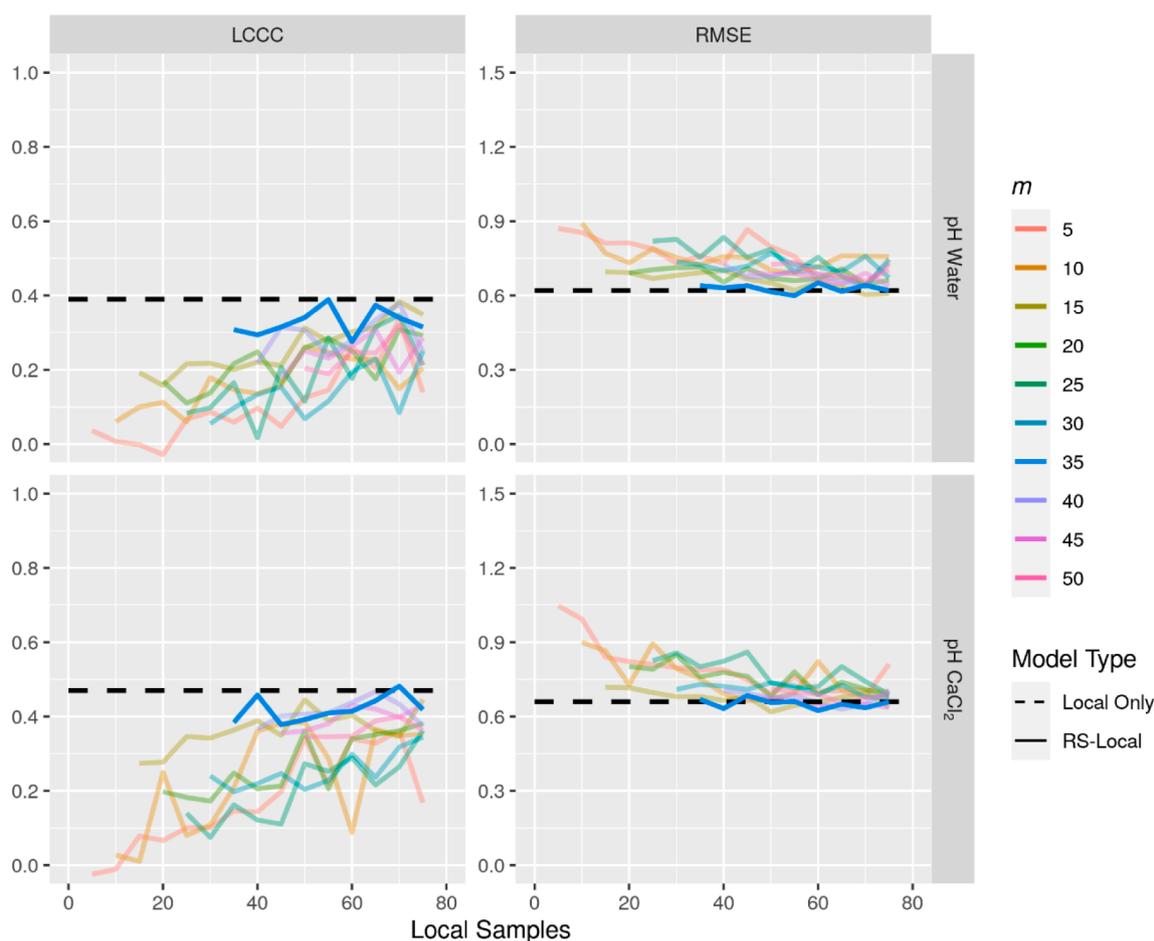


Fig. 11. LCCC and RMSE values for RS-Local selected national and local combined model predictions of pH in water and pH in CaCl₂ generated from the filtered and EPO-transformed CSIRO national Australian Vis-NIR soil spectral library using differing sizes of comparison set m , represented by differing colours and additional local data gathered in field condition. The corresponding goodness of fit metrics from a 50 sample BARS LTT local only model are presented as a dashed black line. Highlighted here are the smallest m sets which, when supplemented with additional spiking samples, achieve LCCCs higher, and RMSE values lower than the 50 sample BARS LTT local only reference model (i.e. for pH in water blue line with m 35 and 20 local samples; and for pH in CaCl₂ blue line with m 35 and 5 local samples).

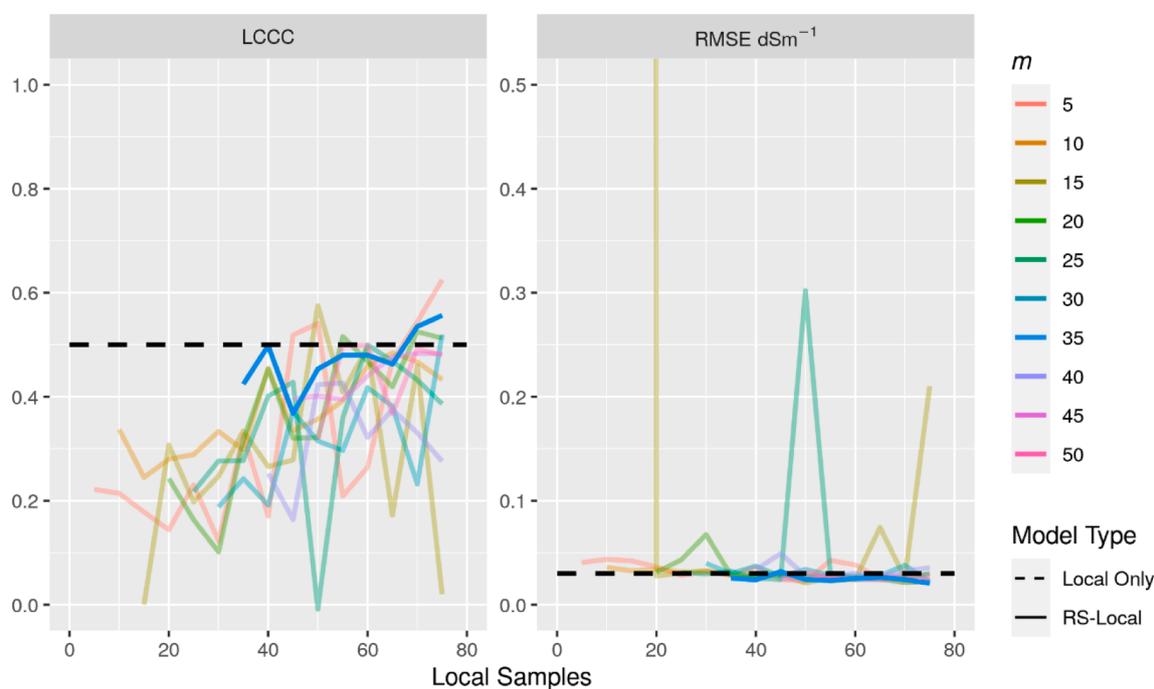


Fig. 12. LCCC and RMSE values for RS-Local selected national and local combined model predictions of EC from the filtered and EPO-transformed CSIRO national Australian Vis-NIR soil spectral library using differing sizes of comparison set m , represented by differing colours and additional local data gathered in field condition. The corresponding goodness of fit metrics from a 50 sample BARS LTT local only model are presented as a dashed black line. Highlighted here is the smallest m set which, when supplemented with additional spiking samples, achieves an LCCC higher, and an RMSE lower than the 50 sample BARS LTT local only Cubist reference model (i.e. blue line with m 35 and 5 local samples).

that of the 50 sample BARS LTT local only model at around 40 local observations. As with pH in CaCl_2 , the first RS-Local selected national and local combined model to approach the prediction quality of the 50 sample BARS LTT local only model consisted of an m set of size 35, and an additional 5 spiking samples (Fig. 12). The RMSE was again lower than the 50 sample BARS LTT local only model, at 0.02 dSm^{-1} and the LCCC was equivalent to that seen in the 50 sample BARS LTT local only model (Table 3). These metrics demonstrate a minor improvement in accuracy to the 50 sample BARS LTT local only model, but greater improvements were seen in the RS-Local selected national and local combined models generated using 45, or 50 local samples, seen in m values of 5 and 15. The model which held the best performance metrics with 50, or fewer samples was a model with an m value of 15, and an additional 35 spiking samples, which held an LCCC of 0.57, and an RMSE of 0.02 dSm^{-1} . The overall model quality of the EC models was much more variable when compared to the SOC, or pH models, with several models with RMSE values greater than 0.2 dSm^{-1} , in the case of the model with an m set of 15 observations, and no additional local spikes, the RMSE value was 28.19 dSm^{-1} .

The RS-Local selection approach for CEC and exchangeable cation predictions on field condition soils using the existing CSIRO national Australian Vis-NIR soil spectral library again did not rapidly exceed the prediction quality of their corresponding 50 sample BARS LTT local only model. For CEC, the first model to approach the quality of the 50 sample BARS LTT local only model used an m of size 10, and an additional 30 samples, achieving an LCCC of 0.79, and an RMSE of $2.34 \text{ cmol} + \text{kg}^{-1}$ (Fig. 13), displaying a slightly poorer fit of observed to predicted than the BARS LTT local only model, at an LCCC of 0.81 (Table 3), and a worse accuracy, against the local RMSE of $2.07 \text{ cmol} + \text{kg}^{-1}$. As with EC, there is a model which leverages 50 local samples, and an selected subset of spectra from the existing CSIRO national Australian Vis-NIR soil spectral library with superior prediction quality to the BARS LTT local only 50 sample model, featuring an m set of size 25, and 25 additional local observations. This model reported an LCCC of 0.83 and an RMSE of $1.94 \text{ cmol} + \text{kg}^{-1}$, showing a strong improvement over the BARS LTT

local only model. As for individual exchangeable cations, for exchangeable Ca, the 50 sample BARS LTT local only model was exceeded at 40 local samples, in a model constructed using an m of 5 samples, and an additional 35 local samples. This model reported an LCCC of 0.69, and an RMSE of $1.52 \text{ cmol} + \text{kg}^{-1}$, exceeding the BARS LTT local only model, but this was the only combination of m sizes and additional samples to achieve this using 50 or fewer local observations. No exchangeable K model surpassed the 50 sample BARS LTT local only model, with the closest model using 50 local observations, a combination of 25 spiking samples, and an m set of 25, achieving an LCCC of just 0.39, lower than the 0.41 of the BARS LTT local only model (Table 3). RS-Local selected national and local combined exchangeable Mg models similarly did not exceed the BARS LTT local only model using 50 or fewer samples. The closest model used 45 local observations, and had an m of 20 samples, and an additional 25 spiking samples, but the LCCC was just 0.69. The exchangeable Na RS-Local selected national and local combined models all held quite poor fits, with none approaching the quality of the 50 sample BARS LTT local only model. Even using as many as 75 local samples, the greatest LCCC achieved was 0.39, well below the 0.49 of the BARS LTT local only model.

3.4. Evaluation of leveraging the CSIRO national Australian Vis-NIR soil spectral library using the memory-based learning spectrum-based learning approach

Spectrum based learning (SBL) selected national and local combined model predictions for the nine studied soil properties also demonstrated a large potential for producing attribute predictions of similar quality to the BARS LTT local only models using equivalent or reduced numbers of laboratory analysed samples. SOC predictions showed LCCC values ranging from 0.57, at a k value of 250, and XYr_5 of 40 samples, to 0.79, with a k of 50, and XYr_5 of 65 (Fig. 14). RMSE values were also similar to the BARS LTT local only model (Table 3), and generally decreased with an increasing number of supplemented local samples, with the lowest RMSE also at a k of 50 and an XYr_5 of 65. Generally, lower k values held

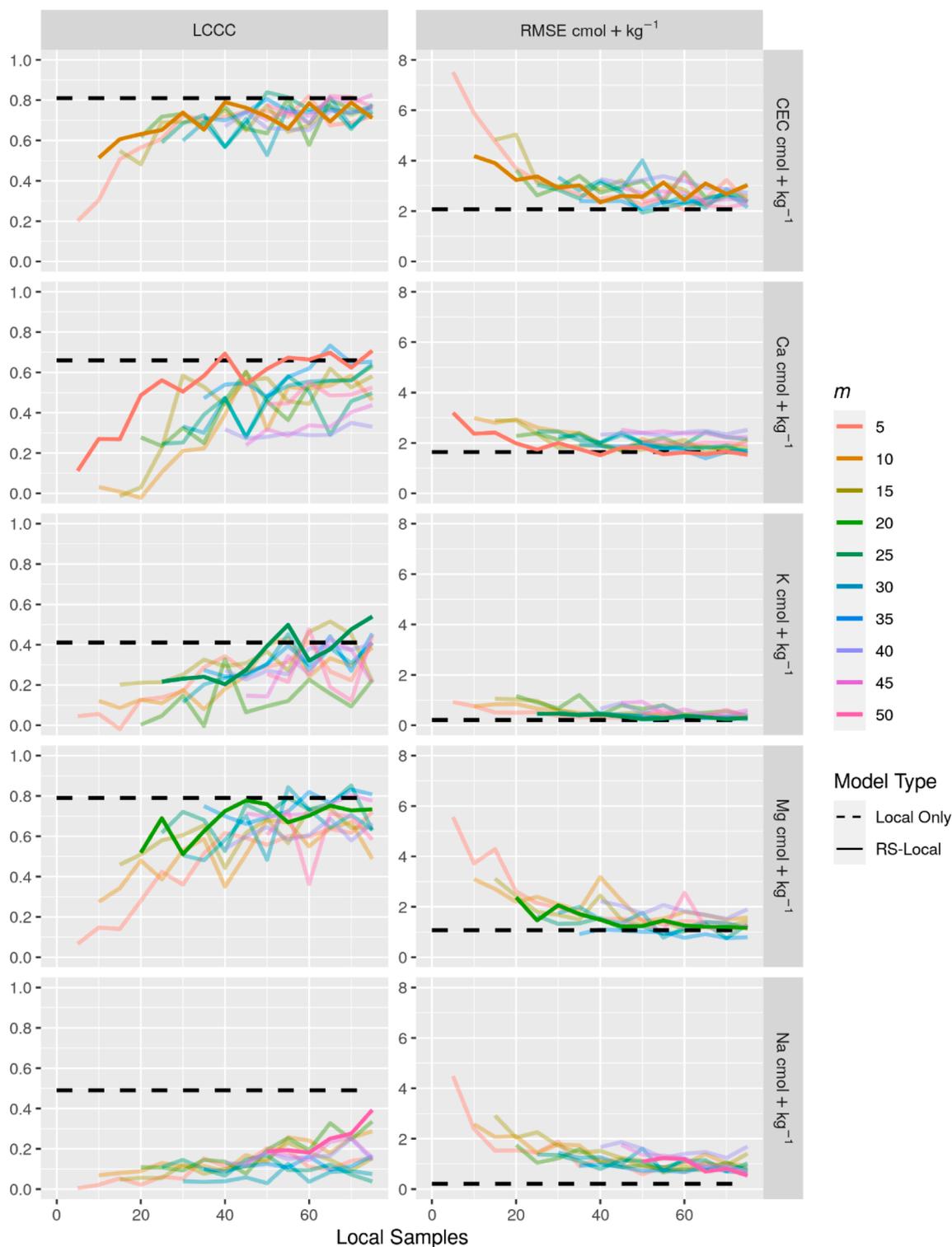


Fig. 13. LCCC and RMSE values for RS-Local selected national and local combined model predictions of CEC and exchangeable Ca, K, Mg, and Na from the filtered and EPO-transformed CSIRO national Australian Vis-NIR soil spectral library using differing sizes of comparison set m , represented by differing colours and additional local data gathered in field condition. The corresponding goodness of fit metrics from a 50 sample BARS LTT local only model are presented as a dashed black line. Highlighted here are the smallest m sets which, when supplemented with additional spiking samples, achieve LCCCs higher, and RMSE values lower than the 50 sample BARS LTT local only reference model (i.e. for CEC brown line with m 10 and 30 local samples; for each Ca red line with m 5 and 35 local samples; for each K dark green line with m 25 and 25 local samples; for each Mg light green line with m 20 and 25 local samples).

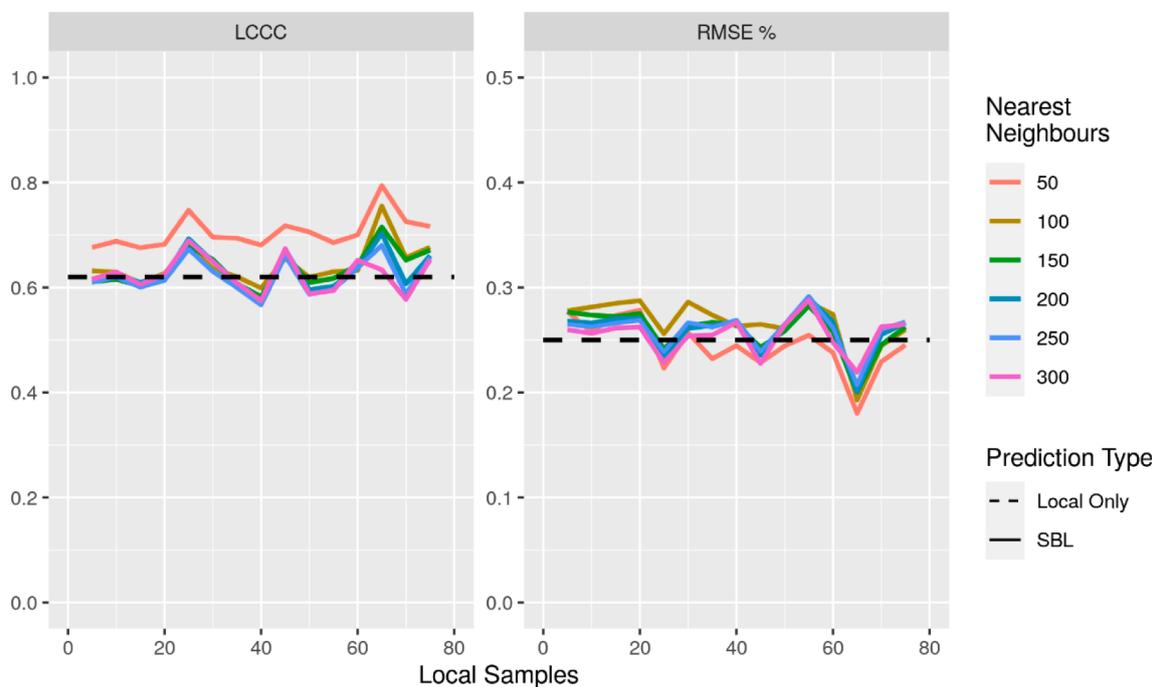


Fig. 14. LCCC and RMSE values for SBL selected national and local combined model predictions for soil organic carbon from the filtered and EPO-transformed CSIRO national Australian Vis-NIR soil spectral library, constructed using a range of spectral nearest neighbours (k) represented by differing colours and supplemented with additional local data gathered in field condition. The corresponding goodness of fit metrics from a 50 sample BARS LTT local only model are presented as a dashed black line.

LCCC values larger than larger values of k , for predictions with the same XYr_s set. All predictions generated with a k of 50 samples exceeded the LCCC of the BARS LTT local only SOC model, however, it must be acknowledged that as the supplemented XYr_s values approached and exceeded the value of k , these models may have approached local only predictions. Spectra collected from a geographically close region, under the same conditions and pre-treatment, may share similar spectral features, and be clumped in a multivariate space, as seen when comparing the local and national datasets used in this study (Fig. 2), leading to primarily local samples being selected through SBL. Regardless, many SBL predictions were of equal quality to the 50 sample BARS LTT local only SOC model, using fewer local observations, most visible at an XYr_s of 25.

Unlike SOC, predictions of pH in water and CaCl_2 did not surpass the quality of the BARS LTT local only models with small values of XYr_s , even with low k values. Only one SBL configuration provided national and local combined model predictions which surpassed the LCCC of a BARS LTT local only model for pH in water, with an XYr_s set 40 with a k value of 50 nearest neighbours (Fig. 15). This configuration did provide an LCCC of 0.43, marginally above the 0.39 of the BARS LTT local only model (Table 3), but the RMSE of the SBL never reached a value lower than the BARS LTT local only model. SBL selected national and local combined pH in CaCl_2 prediction models were of similar quality to the pH in water predictions, with no predictions with an RMSE lower than the BARS LTT local only model and a peak LCCC of 0.42 with a k of 50 and XYr_s of 40, but as the BARS LTT local only predictions of pH in CaCl_2 were superior to those of pH in water, no SBL selected national and local combined model predictions surpassed the quality of the BARS LTT local only model.

SBL selected national and local combined EC model predictions, were not of quality approaching the BARS LTT local only model, even when available local data surpassed the 50 samples used in the local only model. An LCCC of 0.32 and RMSE of 0.04 dSm^{-1} were both optimised in predictions generated with a k of 50, and a XYr_s of 75 (Fig. 16). These predictions are quite poor, but unlike with the RS-Local selection method, RMSE and LCCC values were clustered relatively close together.

Whilst predictions of soil EC by NIR and Vis-NIR spectroscopy are not expected to be of the highest standard (Viscarra Rossel and McBratney, 2008), the low-quality predictions of the national and local combined models selected with both library constraining methods may be enhanced by the highly disordered nature of the EC dataset within the CSIRO national Australian Vis-NIR soil spectral library. Many of the data points either with high spectral similarity selected using SBL or randomly selected for use when identifying a subset through RS-Local may be incomparable to observed values of EC at the BARS LTT, causing large leverage in the models, even post transformations, due to a very high population of outliers. This difficulty in generating a consistent, suitable prediction reinforces the need for these library constraining methods to only be employed for predictions in cases where the reference libraries used to select from are of high quality, and observations held within are trusted.

Predictions of CEC and exchangeable cations generated through SBL selected national and local combined models were, as with RS-Local, generally of quality approaching the BARS LTT local only model. The best performing CEC predictions were seen at a k of 150 and an XYr_s of 65, where a LCCC of 0.85 and an RMSE of $1.93 \text{ cmol} + \text{kg}^{-1}$ were achieved (Fig. 17). No single prediction set below a XYr_s of 50 surpassed both the LCCC (0.81) and RMSE ($2.07 \text{ cmol} + \text{kg}^{-1}$) of the BARS LTT local only predictions (Table 3), but with a k of 150 and XYr_s of 35, produced predictions only marginally poorer, at a LCCC of 0.79, and an RMSE of $2.2 \text{ cmol} + \text{kg}^{-1}$. Predictions for exchangeable Ca using SBL selected national and local combined models were poorer than those seen using RS-Local, with no prediction using 50 or less local spectra producing predictions surpassing the BARS LTT local only model. Prediction qualities began to stabilise around LCCs of 0.52–0.54 at an XYr_s of 20 samples, and RMSE values similarly stabilised between 0.2 and $0.25 \text{ cmol} + \text{kg}^{-1}$. Predictions do not improve until XYr_s values of 65, and did not surpass BARS LTT local only prediction metrics, a LCCC of 0.66 and RMSE of $1.64 \text{ cmol} + \text{kg}^{-1}$. Exchangeable K predictions showed more promising potential for the application of SBL, with predictions surpassing the BARS LTT local only LCCC with 0.41, and approaching the RMSE of $0.21 \text{ cmol} + \text{kg}^{-1}$. The best predictions below

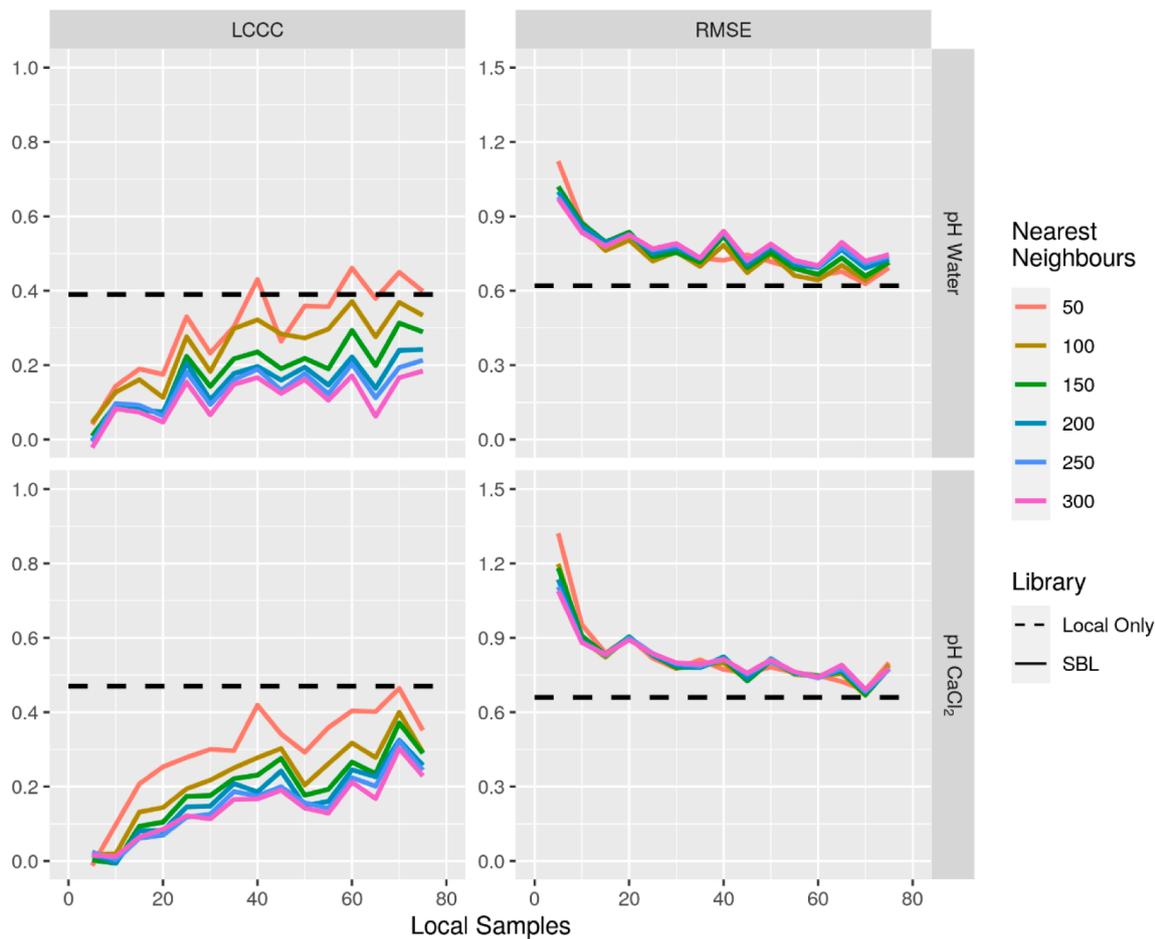


Fig. 15. LCCC and RMSE values for SBL selected national and local combined model predictions for pH in water and pH in CaCl₂ from the filtered and EPO-transformed CSIRO national Australian Vis-NIR soil spectral library, constructed using a range of spectral nearest neighbours (*k*) represented by differing colours and supplemented with additional local data gathered in field condition. The corresponding goodness of fit metrics from a 50 sample BARS LTT local only model are presented as a dashed black line.

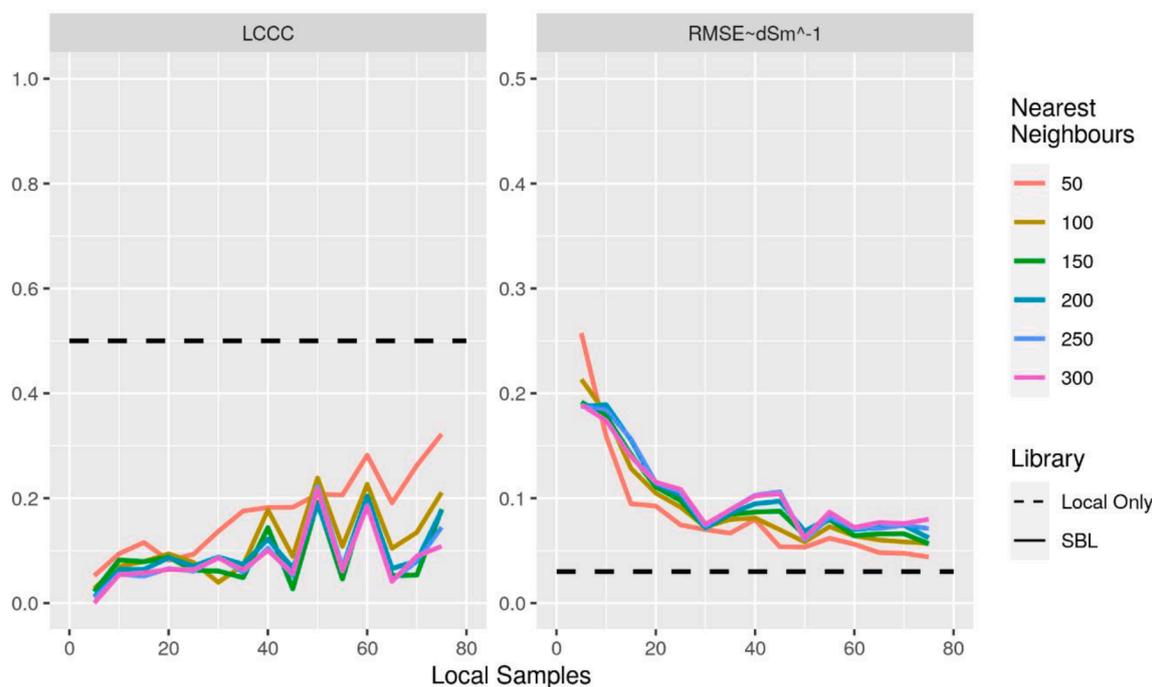


Fig. 16. LCCC and RMSE values for SBL selected national and local combined model predictions for electrical conductivity from the filtered and EPO-transformed CSIRO national Australian Vis-NIR soil spectral library, constructed using a range of spectral nearest neighbours (k) represented by differing colours and supplemented with additional local data gathered in field condition. The corresponding goodness of fit metrics from a 50 sample BARS LTT local only model are presented as a dashed black line.

a XYr_s of 50, were seen at a k of 250, and an XYr_s of 45, where a LCCC of 0.48 and RMSE of $0.22 \text{ cmol} + \text{kg}^{-1}$ were achieved. Exchangeable Mg predictions show the greatest quality of all cations observed under SBL, with a LCCC of 0.89 and RMSE of $0.64 \text{ cmol} + \text{kg}^{-1}$, both improvements over the 50 sample BARS LTT local only model, using a XYr_s of only 30 samples, and predictions built off 250 nearest neighbours. Predictions of exchangeable Na, whilst better than those generated using RS-Local for the national and local combined model selection, were still poor under the SBL approach tested. Most predictions generated did not approach the quality of the BARS LTT local only model, except those with an XYr_s of 65 samples, where predictions using 100 nearest neighbours showed an LCCC of 0.4, and an RMSE $0.232 \text{ cmol} + \text{kg}^{-1}$, approaching the LCCC of 0.49 and RMSE of $0.21 \text{ cmol} + \text{kg}^{-1}$ seen in the BARS LTT local only predictions.

3.5. Advantages, limitations and further research

Both library constraining methods investigated here demonstrate capacity to leverage an existing large spectral library to produce national and local combined model predictions that were of equal quality to a 50 sample BARS LTT local only Cubist model, also using fewer local observations in some scenarios. Predictions of SOC, CEC and exchangeable Ca and Mg were of equal or superior prediction quality using one, or both selection methods. These results indicate that the library constraining methods explored here are suitable tools for selection of laboratory condition spectra from large spectral libraries to apply to field condition spectra, after filtering and the removal of the influence of field conditions on spectral responses through EPO transformation. Furthermore, the observed reduction in the amount of local analytical data required for high quality predictions, can provide large cost savings for soil monitoring programs such as SOC accounting. For this work, if the adapted RS-Local selection approach for the prediction of SOC is employed as an example for cost savings (Fig. 10), at a laboratory analysis cost of up to \$40 AUD per sample, effective employment of the adapted RS-Local implementation can provide a total saving of \$1200

AUD, through reducing laboratory analysis needs to only 20 local samples, compared to the 50 recommended local samples under the Australian Government ERF Soil Carbon Methodology 2021. Having stated this, any spectral library used for soil spectral inference should be appropriate to adequately describe the soil variability of the target area (Viscarra Rossel et al., 2008). As the degree to which a given site is represented within a large spectral library will vary, it is important to test the suitability of these approaches for different agro-ecosystems and different large spectral libraries. As such, it may be unwise to advocate for a reduction in local sampling within monitoring programs such as the Australian Government ERF Soil Carbon Methodology 2021 without first establishing that a site is well covered by data within an existing soil library, and that the relative improvement provided by the library significantly exceeds that provided by further local data collection.

The library constraining methods were not particularly good at selecting useful spectra from the CSIRO national Australian Vis-NIR soil spectral library when the local relationship between a given soil property and spectral response was poor. This was seen in the predictions of pH, EC and exchangeable Na, likely due to the adapted RS-Local method optimising for an already poor local relationship between soil property and spectral response (Table 3), or identifying spectrally similar, but poorly predicting local spectra as nearest neighbours in the case of SBL. EC predictions generated with the adapted RS-Local method demonstrate the need for quality spectral libraries as a prerequisite for library constraining techniques, with poor quality analytical data driving highly stochastic prediction qualities. These poor-quality results reinforce that data curation and cleaning must be an essential component of any workflow intending to use these methods in practical applications.

As for suitability of these methods for use in soil monitoring or environmental incentive programs, like the Australian Government ERF Soil Carbon Methodology 2021, both methods appear to provide a working solution to integrate existing large spectral libraries like the CSIRO national Australian Vis-NIR soil spectral library, but each method holds distinct benefits, and their utility depends on the intended use case and application (Table 4). The adapted RS-Local method produces a

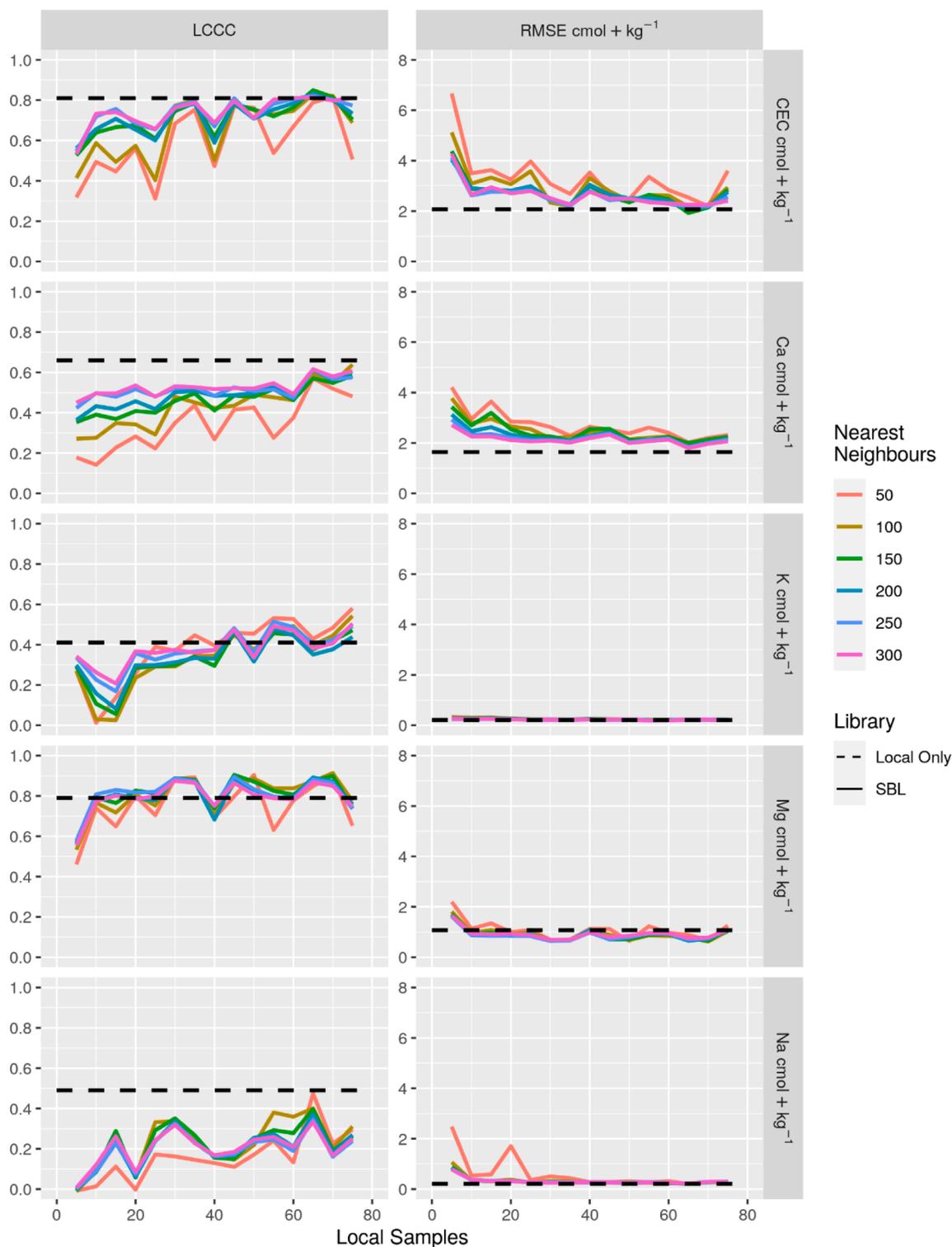


Fig. 17. LCCC and RMSE values for SBL selected national and local combined model predictions for CEC and exchangeable Ca, K, Mg and Na from the filtered and EPO-transformed CSIRO national Australian Vis-NIR soil spectral library, constructed using a range of spectral nearest neighbours (k) represented by differing colours and supplemented with additional local data gathered in field condition. The corresponding goodness of fit metrics from a 50 sample BARS LTT local only model are presented as a dashed black line.

singular model, and localised library, making predictions and datasets traceable and easily repeatable, but the iterations required to run a large number of models to identify that localised library may limit its adoption by smaller scale recipients of incentive programs. Conversely, the applied SBL method merely supplements an existing library with local samples and produces a unique model for each individual target site,

which may hold less traceability than the rigours needed in an incentive programme or stock accounting methodology. However, the implementation of this method is comparatively much simpler, requiring lower computational overhead, making it a much more accessible solution for practical application.

Table 4
Summary of spectral selection methods (Adapted RS-Local vs Spectrum based learning).

Adapted RS-Local	Spectrum based learning
Model can be saved, and reused	Model always dynamic
Selected by spectra and predictor variable	Selected by spectra alone
Large computational overhead to initiate - HPC/Cloud Computing semi-necessary	Simple to initiate, low computation power
Can produce good local soil analyte estimates in situations of low local analytical information	Can produce good local soil analyte estimates in situations of low local analytical information

4. Conclusions

For the first time this research has shown how a large, laboratory grade spectral library, such as the CSIRO national Australian Vis-NIR soil spectral library can be employed in a useful manner to deliver predictions on field condition soils, with low-cost, miniaturised NIR spectrometers. This work demonstrates how condition correction matrices may be developed from soil spectra retrieved from high-resolution spectrometers, and with appropriate filtering to account for differences in resolution, be applied to sensors with lower resolutions and smaller spectral ranges. Moreover, spectra collected in field condition can be transformed to remove undesired environmental influences and be made comparable to existing collected spectra stored in laboratory grade spectral libraries collected in soil laboratory condition (air-dry and < 2 mm). In this manner, datasets stored in laboratory grade large spectral libraries can be made more accessible, and essentially sensor and condition agnostic.

Further, the work shows how large spectral libraries may be localised post filtering and transformation to identify useful predictors at field scale. Existing workflows utilising memory-based learning and similar approaches successfully identified useful sub-sets from the CSIRO national Australian Vis-NIR soil spectral library, providing predictions for some soil properties which exceeded a 50 sample BARS LTT local only model. Ultimately, by enabling the meaningful use of existing soil spectral libraries with low-cost sensors, the barriers to entry to soil monitoring practices can be lowered, both in the up-front costs of standard laboratory grade spectrometers, and the significant analytical costs of establishing local spectral libraries by allowing predictions to be informed by a prior compiled soil spectral library. This can enable better uptake and engagement with soil-focussed environmental incentive programs, to both benefit land holders, and help legislators meet commitments to improve natural capital resources.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: James Moloney reports financial support was provided by Australian Government Department of Climate Change Energy the Environment and Water. Brendan Malone is a member of the editorial advisory board of Geoderma.

Data availability

Data will be made available on request.

Acknowledgements

This research built on project SCMICF000073 "Australia's national soil spectral library empowering rapid SOC measurement" part of the National Soil Carbon Innovation Challenge – Feasibility Study Grants. We thank CSIRO, Hone and the Department of Climate Change, Energy, the

Environment and Water (DCCEEW) and the Australian Government that supported DCCEEW with funding.

This research was supported by resources and expertise provided by CSIRO IMT Scientific Computing, including the high-performance computing facilities from CSIRO that were used for iterating through RS-Local models.

We acknowledge Gerard Ryan for assistance with the soil sampling for this study; Simon Wheeler, Chris Brown, Brenton Gray, Scott Howard, Dan Shiley and Antony Martin from Hone for valuable discussions; and Ross Searle, Gerard Grealish and Mark Glover for their feedback on parts of this research.

James Moloney was supported by a Research Plus CSIRO Early Research Career postdoctoral fellowship (PDF21-1-1045).

References

- Ackerson, J.P., Morgan, C.L.S., Ge, Y., 2017. Penetrometer-mounted VisNIR spectroscopy: Application of EPO-PLS to in situ VisNIR spectra. *Geoderma* 286, 131–138. <https://doi.org/10.1016/j.geoderma.2016.10.018>.
- Barthès, B.G., Kouakoua, E., Coll, P., Clairrotte, M., Moulin, P., Saby, N.P.A., Le Cadre, E., Etayo, A., Chevallier, T., 2020. Improvement in spectral library-based quantification of soil properties using representative spiking and local calibration – The case of soil inorganic carbon prediction by mid-infrared spectroscopy. *Geoderma* 369, 114272. <https://doi.org/10.1016/j.geoderma.2020.114272>.
- Batjes, N. H. 2014. A globally distributed soil spectral library visible near infrared diffuse reflectance spectra.
- Beč, K.B., Grabska, J., Siesler, H.W., Huck, C.W., 2020. Handheld near-infrared spectrometers: Where are we heading? *NIR News* 31 (3–4), 28–35. <https://doi.org/10.1177/0960336020916815>.
- 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 Trends Anal. Chem.* 29 (9), 1073–1081. <https://doi.org/10.1016/j.trac.2010.05.006>.
- Bongiovanni, R., Lowenberg-Deboer, J., 2004. Precision Agriculture and Sustainability. *Precis. Agric.* 5 (4), 359–387. <https://doi.org/10.1023/B:PRAG.0000040806.39604.a>.
- Bornemisza, E., Constenla, M., Alvarado, A., Ortega, E.J., Vasquez, A.J., 1979. Organic carbon determination by the Walkley-Black and dry combustion methods in surface soils and Andeop profiles from Costa Rica. *Soil Sci. Soc. Am. J.* 43 (1), 78–83.
- Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* 132 (3–4), 273–290.
- Dematté, 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., Francolino, M.R., Schaefer, C.E.G.R., Vicente, L.E., dos Santos, U.J., de Sá Barreto 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, E.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., Brighenti, I., de Sousa Mendes, W., Safaneli, 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.
- Environment and Energy. 2018, January 25. Carbon Credits (Carbon Farming Initiative—Measurement of Soil Carbon Sequestration in Agricultural Systems) Methodology Determination 2018 (au). Attorney-General's Department. <https://www.legislation.gov.au/Details/F2018L00089/Html/Text>, <http://www.legislation.gov.au/Details/F2018L00089>.
- European Space Agency. 2021. Copernicus Sentinel-2. MSI Level-2A BOA Reflectance Product. Collection 1. <https://doi.org/10.5270/S2-znk9xsj>.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* 202, 18–27.
- Herrick, J.E., 2000. Soil quality: An indicator of sustainable land management? *Appl. Soil Ecol.* 15 (1), 75–83. [https://doi.org/10.1016/S0929-1393\(00\)00073-1](https://doi.org/10.1016/S0929-1393(00)00073-1).
- Huck, C.W., 2021. New trend in instrumentation of NIR spectroscopy—Miniaturization. In: Ozaki, Y., Huck, C., Tsuchikawa, S., Engelsens, S.B. (Eds.), *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*. Springer, Singapore, pp. 193–210. https://doi.org/10.1007/978-981-15-8648-4_8.
- Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., Vohland, M., 2019. In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the assessment of organic carbon in soils. *Geoderma* 355, 113900. <https://doi.org/10.1016/j.geoderma.2019.113900>.
- Industry, Energy and Emissions Reduction. (2021, December 1). *Carbon Credits (Carbon Farming Initiative—Estimation of Soil Organic Carbon Sequestration Using Measurement and Models) Methodology Determination 2021* (au). Attorney-General's Department.

- <https://www.legislation.gov.au/Details/F2021L01696/Html/Text>, <http://www.legislation.gov.au/Details/F2021L01696>.
- Isbell, R.F., 2021. The Australian Soil Classification. CSIRO Publishing. <https://doi.org/10.1071/9781486314782>.
- Islam, K., Singh, B., McBratney, A., 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Soil Res.* 41 (6), 1101–1114. <https://doi.org/10.1071/sr02137>.
- Johnston, A.J., Pogson, D.J., Thomas, O.D., Watkins, J.J., Glen, R.A., 2013. Boorowa 1: 100 000 Geological Sheet 8629, 1st edition. Geological Survey of New South Wales, Maitland.
- Lin, L.-K., 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45 (1), 255.
- Lobsey, C.R., Viscarra Rossel, R.A., Roudier, P., Hedley, C.B., 2017. Rs-local data-mines information from spectral libraries to improve local calibrations. *Eur. J. Soil Sci.* 68 (6), 6 <https://doi.org/10.1111/ejss.12490>.
- Malone, B., Stockmann, U., Glover, M., McLachlan, G., Engelhardt, S., Tuomi, S., 2022. Digital soil survey and mapping underpinning inherent and dynamic soil attribute condition assessments. *Soil Secur.* 6, 100048 <https://doi.org/10.1016/j.soisec.2022.100048>.
- McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. *Geoderma* 213, 203–213. <https://doi.org/10.1016/j.geoderma.2013.08.013>.
- McKenzie, N., Coughlan, K., Cresswell, H. (Eds.), 2002. *Soil Physical Measurement and Interpretation for Land Evaluation*. CSIRO Publishing.
- Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B.S., Field, D.J., Gimona, A., Hedley, C.B., Hong, S.Y., Mandal, B., Marchant, B.P., Martin, M., McConkey, B.G., Mulder, V.L., O'Rourke, S., Richer-de-Forges, A.C., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C.-C., Vågen, T.-G., van Wesemael, B., Winowiecki, L., 2017. Soil carbon 4 per mille. *Geoderma* 292, 59–86.
- 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.
- 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>.
- Ng, W., Husnain, Anggria, L., Siregar, A.F., Hartatik, W., Sulaeman, Y., Jones, E., Minasny, B., 2020. Developing a soil spectral library using a low-cost NIR spectrometer for precision fertilization in Indonesia. *Geoderma Reg.* 22, e00319. <https://doi.org/10.1016/j.geodrs.2020.e00319>.
- 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., Dematté, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Wetterlind, J., 2015. Chapter Four - Soil spectroscopy: An alternative to wet chemistry for soil monitoring. In: Sparks, D.L. (Ed.), *Advances in Agronomy*, Vol. 132. Academic Press, pp. 139–159. <https://doi.org/10.1016/bbs.agron.2015.02.002>.
- Pannell, D.J., 2008. Public benefits, private benefits, and policy mechanism choice for land-use change for environmental benefits. *Land Econ.* 84 (2), 225–240. <https://doi.org/10.3368/le.84.2.225>.
- Paustian, K., Collier, S., Baldock, J., Burgess, R., Creque, J., DeLonge, M., Dungait, J., Ellert, B., Frank, S., Goddard, T., Govaerts, B., Grundy, M., Henning, M., Izaurrealde, R.C., Madaras, M., McConkey, B., Porzig, E., Rice, C., Searle, R., Seavy, N., Skalsky, R., Mulhern, W., Jahn, M., 2019. Quantifying carbon for agricultural soil management: From the current status toward a global soil information system. *Carbon Manage.* 10 (6), 567–587.
- Quinlan, John R., 1992. Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence, 92.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Dematté, 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, 268–279.
- Ramirez-Lopez, L., Stevens, A., Viscarra Rossel, R., Lobsey, C., Wadoux, A., Breure, T., 2016. Resemble: Regression and similarity evaluation for memory-based learning in spectral chemometrics. *R Package Version 1* (2), 2.
- Rayment, G.E., Lyons, D.J., 2011. *Soil Chemical Methods: Australasia*, Vol. 3. CSIRO Publishing.
- Roger, J.-M., Chauchard, F., Bellon-Maurel, V., 2003. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemom. Intel. Lab. Syst.* 66 (2), 191–204. [https://doi.org/10.1016/S0169-7439\(03\)00051-0](https://doi.org/10.1016/S0169-7439(03)00051-0).
- Rose, D.C., Sutherland, W.J., Parker, C., Lobley, M., Winter, M., Morris, C., Twining, S., Ffoulkes, C., Amano, T., Dicks, L.V., 2016. Decision support tools for agriculture: Towards effective design and delivery. *Agr. Syst.* 149, 165–174.
- Rossel, R.A.V., McBratney, A.B., 2008. Diffuse reflectance spectroscopy as a tool for digital soil mapping. In: Hartemink, A.E., McBratney, A., Mendonça-Santos, M.d.L. (Eds.), *Digital Soil Mapping With Limited Data*. Springer Netherlands, Dordrecht, pp. 165–172.
- Roudier, P., 2011. *chrs: A R package for conditioned Latin hypercube sampling*.
- 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>.
- Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., Viscarra Rossel, R.A., 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Sci. China Earth Sci.* 57 (7), 1671–1680. <https://doi.org/10.1007/s11430-013-4808-x>.
- Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., 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>.
- Soriano-Disla, J.M., Janik, L.J., Allen, D.J., McLaughlin, M.J., 2017. Evaluation of the performance of portable visible-infrared instruments for the prediction of soil properties. *Biosyst. Eng.* 161, 24–36. <https://doi.org/10.1016/j.biosystemseng.2017.06.017>.
- Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Chapter Five—Visible and near infrared spectroscopy in soil science. In: Sparks, D.L. (Ed.), *Advances in Agronomy*, Vol. 107. Academic Press, pp. 163–215. [https://doi.org/10.1016/S0065-2113\(10\)07005-7](https://doi.org/10.1016/S0065-2113(10)07005-7).
- Stockmann, U., Benn, D., Farquharson, R., Gregory, L., Karunaratne, S., Malone, B., Moloney, J., Singh, R., Bundy, A., Harsdorf, J., Macdonald, L., McLachlan, G., Tuomi, S., Weligama, K., Wilson, P., 2022. *Digiscap Future Science Platform project final report: Soil Spectral Analysis Platform*.
- Tang, Y., Jones, E., Minasny, B., 2020. Evaluating low-cost portable near infrared sensors for rapid analysis of soils from South Eastern Australia. *Geoderma Reg.* 20, e00240. <https://doi.org/10.1016/j.geodrs.2019.e00240>.
- van Grieken, M.E., Roebeling, P.C., Bohnet, I.C., Whitten, S.M., Webster, A.J., Poggio, M., Pannell, D., 2019. Adoption of agricultural management for Great Barrier Reef water quality improvement in heterogeneous farming communities. *Agr. Syst.* 170, 1–8. <https://doi.org/10.1016/j.agry.2018.12.003>.
- Viscarra Rossel, R.A., Jeon, Y.S., Odeh, I.O.A., McBratney, A.B., 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. *Soil Res.* 46 (1), 1–16. <https://doi.org/10.1071/SR07099>.
- Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., Rossel, R.A.V., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. *Soil Res.* 53 (8), 845–864. <https://doi.org/10.1071/SR14366>.
- Viscarra Rossel, R.A., Lobsey, C.R., Sharman, C., Flick, P., McLachlan, G., 2017. Novel proximal sensing for monitoring soil organic C stocks and condition. *Environ. Sci. Tech.* 51 (10), 10 <https://doi.org/10.1021/acs.est.7b00889>.
- Viscarra Rossel, R.A., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. *Eur. J. Soil Sci.* 63 (6), 6 <https://doi.org/10.1111/j.1365-2389.2012.01495.x>.
- Wadoux, A.M.J.C., Malone, B., Minasny, B., Fajardo, M., McBratney, A.B., 2021. Soil Spectral Inference with R. *Progress in Soil Science*. Springer, Cham. <https://doi.org/10.1007/978-3-030-64896-1>.
- Waiser, T.H., Morgan, C.L.S., Brown, D.J., Hallmark, C.T., 2007. In Situ Characterization of Soil Clay Content with Visible Near-Infrared Diffuse Reflectance Spectroscopy. *Soil Sci. Soc. Am. J.* 71 (2), 389–396. <https://doi.org/10.2136/sssaj2006.0211>.
- 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), 6.