Earth-Science Reviews 254 (2024) 104797

Contents lists available at ScienceDirect

Earth-Science Reviews

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

An imperative for soil spectroscopic modelling is to think global but fit local with transfer learning

Raphael A. Viscarra Rossel [a,*], Zefang Shen [a], Leonardo Ramirez Lopez [b,c], Thorsten Behrens [e,d], Zhou Shi [f], Johanna Wetterlind [g], Kenneth A. Sudduth [h], Bo Stenberg [g], Cesar Guerrero [i], Asa Gholizadeh [j], Eyal Ben-Dor [k], Mervin St Luce [l], Claudio Orellano [b]

[a] Soil & Landscape Science, School of Molecular & Life Sciences, Faculty of Science & Engineering, Curtin University, GPO Box U1987, Perth, WA 6845, Australia
[b] Data Science Department, BÜCHI Labortechnik AG, Meierseggstrasse 40, 9230 Flawil, Switzerland
[c] Imperial College Business School, Imperial College London, Exhibition Rd, London SW7 2AZ, UK
[d] Swiss Compentence Center for Soils (KOBO), School of Agricultural, Forest and Food Sciences (HAFL), Bern University of Applied Sciences, Länggasse 85, 3052 Zollikofen, Switzerland
[e] Soil and Spatial Data Science, Soilution GbR, Heiligegeiststrasse 13, 06484 Quedlinburg, Germany
[f] Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
[g] Swedish University of Agricultural Sciences, Department of Soil & Environment, PO Box 234, SE-532 23 Skara, Sweden
[h] USDA-ARS Cropping Systems and Water Quality Research Unit, Columbia, MO 65211, USA
[i] Department of Agrochemistry & Environment, Universidad Miguel Hernández de Elche, E-03202 Elche, Alicante, Spain
[j] Department of Soil Science & Soil Protection, Faculty of Agrobiology, Food & Natural Resources, Czech University of Life Sciences Prague, Kamycka 129, Suchdol, Prague 16500, Czech Republic
[k] Department of Geography, Porter School of Environmental & Earth Science Faculty of Exact Science, Tel Aviv University, Israel
[l] Swift Current Research & Development Centre, Agriculture & Agri-Food Canada, 1 Airport Road, PO Box 1030, Swift Current, SK S9H 3X2, Canada

ARTICLE INFO

Keywords:
Soil spectral library
vis–NIR spectra
Localization
Transfer learning
Soil organic carbon
Spectroscopic modelling
Machine learning
Multivariate statistics

ABSTRACT

Soil spectroscopy with machine learning (ML) can estimate soil properties. Extensive soil spectral libraries (SSLs) have been developed for this purpose. However, general models built with those SSLs do not generalize well on new 'unseen' local data. The main reason is the different characteristics of the observations in the SSL and the local data, which cause their conditional and marginal distributions to differ. This makes the modelling of soil properties with spectra challenging. General models developed using large 'global' SSLs offer broad, systematic information on the soil-spectra relationships. However, to accurately generalize in a local situation, they must be adjusted to capture the site-specific characteristics of the local observations. Most current methods for 'localizing' spectroscopic modelling report inconsistent results. An understanding of spectroscopic 'localization' is lacking, and there is no framework to guide further developments. Here, we review current localization methods and propose their reformulation as a transfer learning (TL) undertaking. We then demonstrate the implementation of instance-based TL with RS-LOCAL 2.0 for modelling the soil organic carbon (SOC) content of 12 sites representing fields, farms and regions from 10 countries on the seven continents. The method uses a small number of instances or observations (measured soil property values and corresponding spectra) from the local site to transfer relevant information from a large and diverse global SSL (GSSL 2.0) with more than 50,000 records. We found that with $\leq$ 30 local observations, RS-LOCAL 2.0 produces more accurate and stable estimates of SOC than modelling with only the local data. Using the information in the GSSL 2.0 and reducing the number of samples for laboratory analysis, the method improves the cost-efficiency and practicality of soil spectroscopy. We interpreted the transfer by analysing the data, models, and soil and environmental relationships of the local and the 'transferred' data to gain insight into the approach. Transferring instances from the GSSL 2.0 to the local sites helped to align their conditional and marginal distributions, making the spectra-SOC relationships in the models more robust. Finally, we propose directions for future research. The guiding principle for developing practical and cost-effective spectroscopy should be to think globally but fit locally. By reformulating the localization

* Corresponding author.
E-mail address: r.viscarra-rossel@curtin.edu.au (R.A. Viscarra Rossel).

https://doi.org/10.1016/j.earscirev.2024.104797
Received 14 July 2023; Received in revised form 2 April 2024; Accepted 27 April 2024
Available online 8 May 2024

problem within a TL framework, we hope to have acquainted the soil science community with a set of methodologies that can inspire the development of new, innovative algorithms for soil spectroscopic modelling.

1. Introduction

Soil information is critical for environmental protection, food security, and sustainable development (Bouma, 2019). We need soil data at different scales to assess and monitor changes in soil properties and soil health over time (Lehmann et al., 2020). The need for soil information presents an enormous challenge everywhere, particularly in developing countries (Cook et al., 2008) where soil and land degradation cause hunger and malnutrition and the cost of soil analysis is prohibitively expensive (Viscarra Rossel and Bouma, 2016). Soil spectroscopy can play an integral role in providing this information, and there is much international interest in the technology (Viscarra Rossel et al., 2022).

Reflectance spectroscopy is a powerful soil analytical method that relies on the interaction of electromagnetic radiation at specific frequencies, usually in the visible (vis, 400–700 nm), near infrared (NIR, 700–2500 nm) or mid infrared (MIR, 2500–25,000 nm), with the soil constituents. This fundamental physical process provides insights into soil composition and enables the estimation of soil properties (Soriano-Disla et al., 2014). When MIR or NIR energies are emitted onto the soil, the light scatters within the sample, causing the bonds in the molecules present to vibrate and absorb some of that light. The rest is only diffusely returned to a detector, which records the response as a function of wavelength or wavenumber. The fundamental vibrations of bonds in molecules occur in the MIR, and the vibrations can be of different types, e.g., symmetric, asymmetric, bending, scissoring (Griffiths, 2010). These fundamentals cause the excitation of the vibrational modes in the bonds of molecules from their lowest ground energy state to their first excited state. Overtones and combination vibrations occur when the energy transitions are from the ground state to the second or higher vibrational state and when the transitions are between two or more vibrational modes, respectively. NIR spectra result only from overtones and combination vibrations. The process differs in the visible range, where the energies are higher, and spectra result from the excitation of electrons and electronic transitions (Picollo et al., 2019).

Depending on the resolution of the sensor, a single soil vis–NIR or MIR spectrum consists of hundreds to thousands of frequencies, and depending on the spectral range, those frequencies can hold information on soil colour, the iron oxides (e.g., hematite, goethite), clay minerals (e.g., gibbsite, kaolinite, illite, smectite), carbonates and gypsum when they are present in the soil, the types of organic matter, the content of (adsorbed and free) water, and the particle size (Clark et al., 1990; Ben-Dor and Banin, 1995; Nguyen et al., 1991; Viscarra Rossel and Hicks, 2015). Thus, in a single measure, a soil spectrum characterizes the soil's fundamental and multivariate composition, which determines soil properties and functions. The technique is non-destructive, rapid, inexpensive, and precise, making it an indispensable tool for soil analysis.

The absorption at specific wavelengths can be used directly to derive measures of soil colour (Viscarra Rossel et al., 2009), the abundance of iron oxides (Viscarra Rossel et al., 2010), clay minerals (Viscarra Rossel, 2011), and water (Baumann et al., 2022). However, to derive estimates of other soil properties, one must first develop a soil spectral library (SSL) and then model (or calibrate) those soil properties with the spectra. An SSL consists of data pairs with laboratory-measured soil properties and their corresponding spectra. These libraries serve as valuable repositories of soil information because, as explained above, they capture the unique composition of the soil samples and the spectral characteristics associated with specific soil properties. Ideally, SSLs should be developed by design considering the domain of application, the sampling strategy, the soil analytical methods, the spectral range and the protocols used. Often, however, SSLs are built using legacy soil samples stored in archives derived from experiments with different aims and using different analytical methods and spectroscopic protocols (Nocita et al., 2015). Although this is a cost-effective approach for developing SSLs and an excellent way to use archived soils, when developing spectroscopic calibrations, one must carefully consider the quality of the analytical data and the applicability of the models in different domains. There are now many examples of SSLs developed for different regions, countries, continents, and the world (e.g. Shepherd and Walsh, 2002; Viscarra Rossel and Webster, 2012; Stevens et al., 2013; Shi et al., 2015; Viscarra Rossel et al., 2016; Dematte et al., 2019).

The primary aim of soil spectroscopic modelling is to relate soil properties with the information contained in the various frequencies of the spectra to then be able to estimate those soil properties by inputting newly measured soil spectra into the models. Measuring soil spectra is easier, faster, and less expensive than measuring soil properties with conventional methods of soil analysis. Another advantage of the spectroscopic approach is that using the same spectra in the library, one can derive models to estimate many soil properties, of course, as long as corresponding measurements of those soil properties are present in the library. However, the technology is not a panacea for all our soil measurement needs. When the soil's physical, chemical, and biological properties derive from or are associated with the soil's mineral–organic matrix, spectroscopy can reasonably accurately estimate the concentrations of those properties. However, when the constituents are not properties of the soil matrix, the correlations will be weak and only transient at best (Viscarra Rossel et al., 2022).

Much of the earlier research and development in soil spectroscopic modelling relied on multivariate calibrations with principal component regression (PCR) (e.g. Chang et al., 2001) and partial least squares regression (PLSR) (Martens and Næs, 1989). These methods are robust and perform well when the response-spectra relationship is linear, which is more likely when the SSL represents a small, local domain. With the development of larger, more diverse and complex SSLs and the advent of machine learning (ML) in soil science, researchers began testing other methods that can cope better with more extensive, non-linear datasets, for example, using wavelets, random forests (RF), support vector machines (SVM), regression trees, the Gaussian pyramid scale (Viscarra Rossel and Lark, 2009; Viscarra Rossel and Behrens, 2010; Behrens et al., 2022; Vohland et al., 2016). Most recently, coinciding with developments in deep neural networks, studies have also tested these methods for modelling soil properties with spectra (e.g. Liu et al., 2018; Padarian et al., 2019). Although the deep learning algorithms require more training data, are more complex, and their implementation more computationally expensive, they offer some advantages over conventional ML. For example, they can provide automatic pre-processing and extraction of useful feature representations, which streamline the modelling and improve their performance with large data sets (Tsakiridis et al., 2020; Shen and Viscarra Rossel, 2021).

Research on soil spectroscopic modelling and the calibration and validation of predictive functions to estimate soil properties has, over the past decades, helped to establish the value of SSLs and the potential of soil spectroscopy for accurate and cost-effective estimation of soil properties (Li et al., 2022). But, despite the success of such research and the models built using SSLs, their ability to generalize well locally is limited (Shen et al., 2022). Soil properties exhibit significant variability over different spatial scales because they are affected by the local factors that affect soil formation, e.g., climate, organisms, relief, parent material, land management practices, and time (Jenny, 1941). Therefore, models trained on large and diverse SSLs often fail to capture the site-specific soil variation needed for accurate local estimation (Viscarra Rossel et al., 2022).

General models built using entire SSLs, often called 'global' models, offer broad, systematic information on the spectra-soil relationships. The models need to be adjusted and fine-tuned to capture the site-specific characteristics of soil variation accurately and generalize effectively in a local context. Hence, researchers have developed spectral localization techniques that attempt to update local models with information from the large SSLs. These techniques aim to enhance model performance by tailoring the models to the specific characteristics of individual sites, producing accurate local estimates of soil properties. There has been some research to develop such techniques, but they have shown variable success, and research is slow but ongoing. We describe those in Section 2.

The localization of spectroscopic models is a problem that transfer learning (TL) can help to address. TL describes mathematical techniques that leverage information from one source domain to improve performance in another related target (or local) domain. TL is inspired by a human's ability to build upon pre-existing knowledge to solve a new problem faster and more effectively instead of starting from scratch. It aims to transfer helpful and relevant information from a source domain to a local domain, where data may be scarce or unavailable or when the estimation must be rapid. As such, combined with ML and artificial intelligence (AI), TL provides an intuitive framework for soil spectroscopic modelling and a powerful combination for future research and development of practical, deployable soil spectroscopy.

Therefore, our aims are to:

- Describe the localization problem and the current methods for localizing soil spectroscopic models.
- Describe how TL offers a framework for explicitly describing the problem and developing innovative new research and solutions.
- Demonstrate how a TL algorithm can relay helpful information from a large global SSL to 12 local sites worldwide from 10 countries and The Ross Dependency in the seven continents.
- Delve into the results from the TL algorithm to gain insight into the transfer and better understand and interpret it from statistical and scientific perspectives. and
- Propose directions for future research.

2. Localization of spectroscopic modelling

Large, country, continental, and global SSLs have been developed to enable the estimation of a range of soil properties (Viscarra Rossel et al., 2016; Stevens et al., 2013; Dematte et al., 2019; Shi et al., 2015; Viscarra Rossel and Webster, 2012). We can generally find good statistical relationships between soil properties, e.g., soil organic carbon (SOC) and spectra (Viscarra Rossel and Behrens, 2010). However, these empirical models can vary, and when derived with all of the data in a large SSL (i.e., a 'global' model), they often fail to capture accurately the local characteristics of the soil properties at the site (e.g., a field in a farm), particularly when using linear multivariate methods such as PLSR.

Non-linear methods and ML algorithms (Viscarra Rossel and

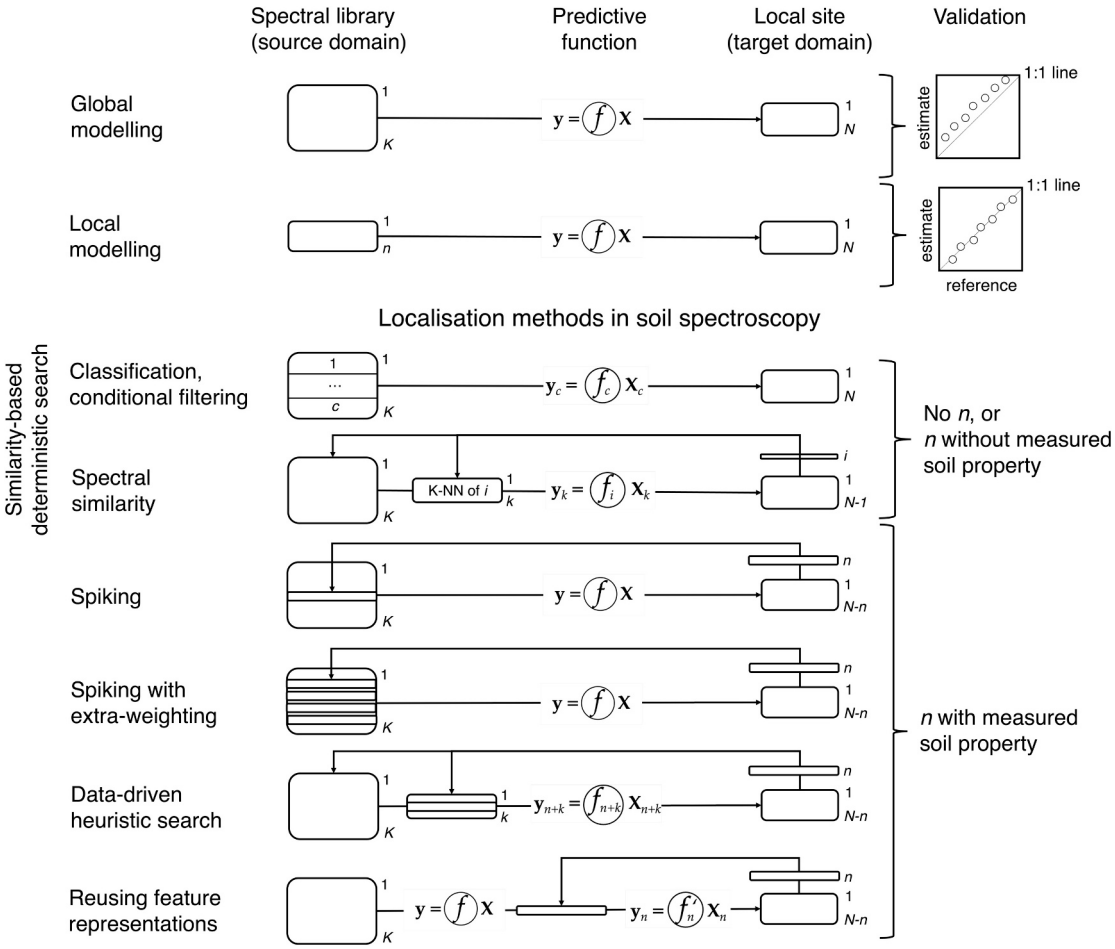

Fig. 1. Summary of methods for localizing soil spectroscopic modelling. Models built using all observations in a soil spectral library (SSL; Global modelling) tend to be biased. Models built using only local observations (Local modelling) are accurate but expensive. Localization methods aim to improve the accuracy and cost-effectiveness of soil spectroscopy. **X** represents the spectral matrix, **y** the soil property vector, f the predictive function; K is the number of observations in the SSL, c and k are subsets of 'localized' observations from the SSL, N is the number of observations in the local data, n is the number of representative local observations with the measured soil property, and i is the i^{th} sample in N.

Behrens, 2010; Shen and Viscarra Rossel, 2021) have helped to make use of large SSLs for analyses and interpretation over country or larger scales, e.g., for continental scale digital soil mapping (Viscarra Rossel et al., 2014, 2019). However, it is now well-understood that even if one accounts for nonlinearities in the data, e.g., by partitioning the dataset, 'global' models tend to not generalize well locally, with spectra from fields or farms (Fig. 1). The reasons might be the complex composition of soil and the diversity of the organic-mineral matrix (Stenberg et al., 2010), which is pertinent when modelling SOC because in many parts of the world, SOC concentrations are generally small (Köchy et al., 2015). Typically, there is an increase in average prediction errors with increasing variability in the 'global' model (e.g. Sankey et al., 2008; Guerrero et al., 2016; Lobsey et al., 2017; Shen et al., 2022). Local models predict well (Fig. 1); however, they incur a greater soil analytical cost and are inefficient because local modelling does not use the information in large and diverse SSLs.

Various approaches have been proposed to better use large SSLs for local predictions of soil properties. Most aim to reduce, minimize or even remove the need for conventional soil analyses and maximize the overall accuracy of the spectroscopic estimates at the local site. Current methods for such localization (Fig. 1) are based on either classification of the SSLs to constrain the modelling, data augmentation, deterministic local search algorithms based on spectral or sample similarities, data-driven stochastic search methods, or the reuse of transformed spectral features, or representations. In Section 2, we review the localization methods in soil spectroscopy, and in Section 3 describe this challenge using a TL framework.Section 3

2.1. Similarity-based deterministic methods

Similarity-based methods attempt to divide large SSLs into smaller groups where linear models can describe the relationships between the soil property and spectra. All use deterministic similarity methods, and not all methods require the measured soil property (Fig. 1).

The most intuitive way to use large, diverse SSLs is to classify them into smaller, more homogeneous subsets with soil samples that share similar characteristics. In this way, the intra-subset variability is smaller than the overall variability of the SSL. One could constrain the SSL with ancillary information that helps to characterize the pedological context and the relationships between soil properties and the spectra, for example, using taxonomic, geographic or land use information (e.g., Sankey et al., 2008; Vasques et al., 2010; Xu et al., 2016; Moura-Bueno et al., 2020). Spectroscopic models can then be developed using data from each subset (Fig. 1). To estimate soil properties at a local site, one uses the model that best captures the characteristics of the local conditions. Although such classification might result in smaller libraries, they might not capture the spectra-response relationship needed for local estimation (e.g., on data from individual fields).

More sophisticated methods that either 'memorize' the training data to find similarities, or that perform data-driven classifications of the spectra include memory-based learning (MBL), local regressions (e.g. Rabenarivo et al., 2013), the spectrum-based learner (SBL) (Ramirez-Lopez et al., 2013), and cubist (e.g. Viscarra Rossel and Webster, 2012).

One of the most common recent methods for similarity-based local search is MBL, *a.k.a k*-nearest neighbours. The MBL methods extract spectrally similar samples (e.g., using the Mahalanobis distance) from the SSL for each observation in the local set, develop a specific calibration with the selected neighbours and predict the unknown local data, effectively deriving a site-specific, local calibration (Fig. 1). The LOCAL (Shenk et al., 1997) and locally weighted regression (LWR) algorithms (Naes et al., 1990; Gupta et al., 2018) and their variants are MBL examples. In LWR, the selected calibration samples are weighted according to the spectral similarity between the SSL and the unknowns. In the SBL, the nearest neighbours from an SSL are selected using distance metrics calculated in principal component space. The training data set for the spectroscopic modelling uses the selected neighbours and the matrix of distances to the unknown samples. Tsakiridis et al. (2020) used SBL to select the nearest neighbours in a continental SSL and used their prediction errors to correct the estimates of soil properties in the test set. The method effectively accounts for non-linear relationships in large and complex data sets since the relationships can be well-described by simple linear models within the neighbourhoods (Ramirez-Lopez et al., 2013). However, the method is computationally expensive for very large SSLs since every new prediction requires the calculation of its similarity to every spectrum in the SSL. The cubist algorithm is a tree-based method that generates rules with linear models at each leaf (Quinlan, 1992). These rules are used to classify the spectra. A new observation is predicted by classifying it and applying the corresponding model. The algorithm, like other regression-tree methods, can accommodate other ancillary data to generate the rulesets. Cubist has been extensively reported in the literature and shown to produce accurate and interpretable models (e.g. Viscarra Rossel and Webster (2012)).

2.2. Spiking and spiking with extra weighting

Simple spiking uses several local observations (response variable with their spectra) to augment a larger SSL before modelling (Fig. 1). The reported success of the approach is mixed. Some studies show that the method can produce more accurate (less biased) estimates of soil properties compared to global models derived with only the SSL, while others report little or no improvement (e.g. Brown, 2007; Sankey et al., 2008; Viscarra Rossel et al., 2009; Guerrero et al., 2010; Wetterlind and Stenberg, 2010; Gogé et al., 2014; Barthès et al., 2020). Guerrero et al. (2014) proposed that spiking could be improved by using multiple copies of the local samples to augment the SSL (Fig. 1). They called this method spiking with extra weighting and proposed that it was better than simple spiking, particularly with larger SSLs, because it improves the leverage of the local data in the models. Both simple spiking and spiking with extra weighting increase the size of the calibration set (Fig. 1). Barthès et al. (2020) showed that the spiking with extra weighting could improve the accuracy of models over simple spiking for estimating soil inorganic carbon. The success of these approaches may be inversely related to the size and diversity of the SSL and the degree of similarity between the SSL, the spiking subset, and the local sites for estimation (Guerrero et al., 2016). The methods are less effective and can fail when the distributions of the SSL and the local data are too dissimilar. That is when the data and model are not all relevant. For example, in Seidel et al. (2019), the SSL was from an entire country, Germany, the local data were from an agricultural field, and the spiking subset had $\leq$30 local observations. In this case, spiking was less accurate than local modelling.

2.3. Data-driven heuristic search

rs-local, a data-driven heuristic search method, was developed by Lobsey et al. (2017). It uses a small number of local observations, or instances (measured soil property and spectra) to select a subset of data from an existing SSL for modelling (Fig. 1). It uses a stochastic selection procedure that repeatedly samples the observations from the SSL without replacement. rs-local does not assume specific relationships between the response variable and the spectra in the SSL (e.g., does not assume linearity); instead, the relationships are specific to the local site. Only the instances from the SSL that perform well on the local data are selected. Thus, the approach filters out data that, when added to linear models consistently increase the inaccuracy of the local predictions. The filtered-out data are inconsistent with the local spectra-response relationship, that may result from erroneous spectra, measurements with different spectrometers, or inaccurate analytical measurements in the SSL. The approach has been shown to produce robust localized vis–NIR and MIR models developed using only a few well selected samples from the local site (Shen et al., 2022; Lobsey et al., 2017; Baumann et al., 2021; Helfenstein et al., 2021). Lobsey et al. (2017) showed that rs-local

performs better than other methods such as 'spiking' and MBL. In Shen et al. (2022), we developed an improved version of RS-LOCAL, which we refer to as RS-LOCAL 2.0. It includes parallelization and is computationally more efficient because parts of the algorithm are implemented in C^{++}.

2.4. Reuse of representations

As we alluded to earlier, in the context of soil spectroscopic modelling, representations are transformations of the input spectra into more informative and compact forms that facilitate localization and modelling (Fig. 1). They can be obtained using a different techniques, depending on the problem (e.g., dimensionality reduction, feature extraction, deep neural networks).

The most common examples are feature extraction or fine-tuning of artificial neural networks, where an existing pre-trained model is updated on new data. Recent studies have explored reusing representations learned in large-scale convolutional neural networks (CNNs) built on continental and global SSLs to improve the local estimation of soil properties (Liu et al., 2018; Padarian et al., 2019; Shen et al., 2022). The approach is based on the notion that initial layers in a CNN learn generic representations while those in latter layers are task-specific (Zeiler and Fergus, 2014; Yosinski et al., 2014). These generic representations, learnt from a large data set, can be reused to improve the accuracy of local modelling. Hence, one fixes the initial layers in a large-scale CNN to implement the method and retrains the remaining layers using the local data. Reusing representations performed well when localizing spectroscopic models from continental to country scales (Padarian et al., 2019). Performance for local estimation, however, varies (Shen et al., 2022). We need more research to improve the method's robustness for local modelling, e.g., by enhancing or selecting only the most relevant representations. Reusing representations in other models is also possible. Ng et al. (2022) extracted representations by training a PLSR on a regional SSL and reusing its loadings on the local spectra. The approach did not consistently improve the estimates compared to local modelling.

2.5. Hybrid methods

Combining the above techniques to extract useful information from a large SSL is also possible. Wetterlind and Stenberg (2010) used spectral neighbours with spiking of a national SSL to improve the local estimation of clay and SOC at four different farms. Shi et al. (2015) proposed using spectral similarities and geographical constraints to model SOC and estimate it locally. They reported improvements in the accuracy of estimates when the SSL was constrained to the geographical region from which the unknown samples originated. The GLOBAL-LOCAL algorithm integrates distance-based spectral similarity and heuristic search methods (St. Luce et al., 2022). It generates SSL subsets containing neighbours of the representative local observations using the Mahalonobis distance on the spectral principal components and develops PLSRs that are evaluated on the representative subset of the local data to select the one that performs best. The method was tested using two relatively homogenous data sets and requires further testing. In a previous study (Shen et al., 2022) combined the reuse of representations and heuristic search with RS-LOCAL 2.0. They trained a CNN on a global SSL and partially retrained the CNN on data selected by RS-LOCAL 2.0. Combining the reused representations with RS-LOCAL 2.0 did not consistently improve the SOC estimation in all the local sites tested.

3. Transfer learning

Implementing TL is like taking what one has learned from an experience and applying it to a different but related situation, making the 'learning' faster, cheaper, easier, or all combined. TL is beneficial when there are only few data available to train a model or when a model could benefit from general domain information to help it capture important patterns and features that are relevant to the new situation, or when model training needs to be fast, accurate, and adaptable to changing requirements and data distributions. TL provides an appealing framework and a set of methodologies that can be applied in the various fields of science, engineering, and ML.

TL is not a new concept. The first report that describes TL was published in the 1970s for pattern recognition using neural networks (Bozinovski, 2020), and research continued throughout the 1980s and 1990s (e.g. Pratt et al., 1991). Since then, TL has attracted increasingly more attention and under different names and related methods, such as 'knowledge transfer', 'inductive transfer', 'meta-learning' and 'multitask learning'. In 2005 the Defense Advanced Research Projects Agency (DARPA) of the United States Department of Defense defined TL as 'the ability of a system to recognize and apply knowledge and skills learned in previous tasks to novel tasks' (Pan and Yang, 2010).

The development and adoption of TL have been driven by several shortcomings of traditional ML, including their need for large volumes of data to train models, the computationally expensive training times, the poor generalization of the 'global' models on 'unseen' local data and the distribution mismatch between the datasets. Advances in information and communications technologies, Graphics Processing Units (GPUs), cloud computing, and the simultaneous and astonishing leaps in the development of AI and ML have also fueled developments in TL. TL remains an active and vibrant area of research, and there have been several publications that report on advances in the techniques and their application (e.g. Pan and Yang, 2010; Weiss et al., 2016; Zhuang et al., 2020; Niu et al., 2020).

3.1. Definition of transfer learning in soil spectroscopic modelling

We must first introduce some terminology, definitions, and notation to understand TL. To enhance its relevance to soil spectroscopic modelling, we have adapted the terminology, definitions and notation from different texts (e.g. Pan and Yang, 2010; Weiss et al., 2016; Zhuang et al., 2020; Niu et al., 2020).

A spectral domain, $\mathscr{D}$ is defined as having a feature (or spectral) space $\mathscr{X}$ and a soil property space $\mathscr{Y}$. $\mathscr{X}$ contains the entire collection of spectra in a matrix $\mathbf{X}$, which has m observations, n features (or spectral intensities at specific frequencies), and a marginal probability distribution $P(\mathbf{X})$. $\mathbf{X}$ holds the spectra as vectors, $\mathbf{x}$, where $\mathbf{x} = \{x_1, \ldots, x_n\} \in \mathscr{X}$ and $x_1, \ldots, x_n$ are the spectral intensities associated with a particular observation. $\mathscr{Y}$ contains the soil properties in a matrix $\mathbf{Y}$ with m observations and p soil properties and with conditional distribution $P(\mathbf{Y}|\mathbf{X})$. Thus, $\mathbf{Y}$ holds the soil property vectors, $\mathbf{y}$, where $\mathbf{y} = \left\{y_1, y_2, \ldots, y_p\right\} \in \mathscr{Y}$, and $y_1, \ldots, y_p$ are the soil property values associated with a particular observation.

Then, given a particular $\mathscr{D}$, its task, $\mathscr{T} = \{\mathbf{X}, \mathbf{y}, f(\cdot)\}$, consists of the training data $\mathbf{X}$, $\mathbf{y}$ (note that in this case, we are training a single soil property, e.g. SOC) and the predictive function $f(\cdot)$ that is not known but 'learned' from the training data. The function $f(\cdot)$ can be a statistical or ML model (e.g., PLSR, CUBIST) or a deep neural network. If the latter, the method is called deep transfer learning (DTL) (Tan et al., 2018). The task can also be represented probabilistically (Weiss et al., 2016), as $\mathscr{T} = \{\mathbf{X}, \mathbf{y}, P(\mathbf{y}|\mathbf{X})\}$, where $P(\mathbf{y}|\mathbf{X})$ is the conditional probability distribution of the soil property given the spectra.

Therefore, given a *source* spectral domain, $\mathscr{D}_s$, with its corresponding learning task $\mathscr{T}_s$, and a *target (or local)* spectral domain, $\mathscr{D}_l$, with its learning task $\mathscr{T}_l$, TL aims to improve the local predictive function, $f_l(\cdot)$ using the relevant and useful information gained from $\mathscr{D}_s$ and $\mathscr{T}_s$. Usually, $\mathscr{D}_s >> \mathscr{D}_l$ and from the definition, the source and local spectral domains and tasks are different, i.e., $\mathscr{D}_s \neq \mathscr{D}_l$ and $\mathscr{T}_s \neq \mathscr{T}_l$, but somewhat related. Thus, four possible TL scenarios emerge (Pan and Yang, 2010; Weiss et al., 2016):

1. The source and local spectral spaces are different, i.e., $\mathscr{X}_s \neq \mathscr{X}_l$. For example, when the source spectra are one type (e.g., MIR) and local spectra another (e.g., vis–NIR).
2. The marginal distributions of the source and local spectra are different, i.e. $P(\mathbf{X}_s) \neq P(\mathbf{X}_l)$. For example, when the source and local spectra are measured with different spectrometer types that measure the same type of spectra. In this case, the types of spectra are the same and share the same feature space (i.e. $\mathscr{X}_l = \mathscr{X}_s$), but their spectral intensities are different. This scenario is often referred to as domain adaptation (e.g., Pan et al., 2011).
3. The source and local soil property spaces are different, i.e. $\mathscr{Y}_s \neq \mathscr{Y}_l$. For example, when the soil property in the source spectral domain is SOC content and that in the local spectral domain is soil organic matter (SOM) content, or a proxy, e.g., soil colour.
4. The conditional distributions of the source and local soil property are different, i.e., $P(\mathbf{Y_s}|\mathbf{X_s}) \neq P(\mathbf{Y_l}|\mathbf{X_l})$. For example, when the soil properties in the source and local data are measured using different methods, in different laboratories, and by different practitioners. In this case, the conditional distributions of the source and local domains are likely to be different because of the variations in analytical procedures.

Of these four scenarios, 2. and 4. are likely to be the most useful and practically in soil spectroscopic TL.

3.2. Categorization of transfer learning for soil spectroscopic modelling

The terminology and definitions used to categorize the different TL scenarios and solutions are somewhat inconsistent in the literature. Possibly due to the considerable research interest and rapidly evolving concepts, algorithms, and applications (e.g. Weiss et al., 2016; Niu et al., 2020; Zhuang et al., 2020).

In their paper, Pan and Yang (2010) classified TL according to the similarities between the source and target (local) domains and the availability of measured response variable values (or 'labels' as referred to in the ML literature). Thus, they proposed three categories: 'inductive', 'transductive', and 'unsupervised' TL—we describe these below. Weiss et al. (2016) used a broader classification that considers TL according to the similarities between the domains regardless of the availability or unavailability of measured response data. They proposed 'homogeneous' and 'heterogeneous' TL. Homogeneous TL occurs when the variable spaces in the source and target domains are similar or closely related, i.e., when $\mathscr{X}_s = \mathscr{X}_l$ and $\mathscr{Y}_s = \mathscr{Y}_l$, and 'heterogeneous' TL occurs when they are dissimilar, i.e., when $\mathscr{X}_s \neq \mathscr{X}_l$ or $\mathscr{Y}_s \neq \mathscr{Y}_l$. Zhuang et al. (2020) reviewed homogeneous TL, and Day and Khoshgoftaar (2017) reviewed heterogeneous TL and the methodologies used for such cross-domain learning.

Here, we borrow from those classifications, and Fig. 2 shows our proposed categorization, which is relevant to our context: soil spectroscopic modelling.

3.2.1. Defining the transfer learnign problem

When implementing TL in soil spectroscopy, the first decision is to assess the similarity between the source and local domains and determine if the problem is homogeneous or heterogeneous (Fig. 2). Since in heterogeneous TL, the feature or response variable spaces are not equivalent, the first aim of their solutions is to reduce the dissimilarity between the spaces so that the problem becomes a homogeneous TL problem. No examples of heterogeneous TL exist in the soil or spectroscopic literature, but exploring the methods may be helpful. Interested readers can find a detailed description of the heterogeneous TL problem and methodologies used to derive solutions in Day and Khoshgoftaar (2017). Homogeneous TL methods aim to reduce the dissimilarities in the marginal and conditional distributions (or both) of the source and local domains. All of the spectroscopic localization methods (Fig. 1) are homogeneous TL problems (Fig. 2). Homogenous and heterogeneous TL methods can be 'inductive', 'transductive', or 'unsupervised' (Fig. 2). We describe these below.

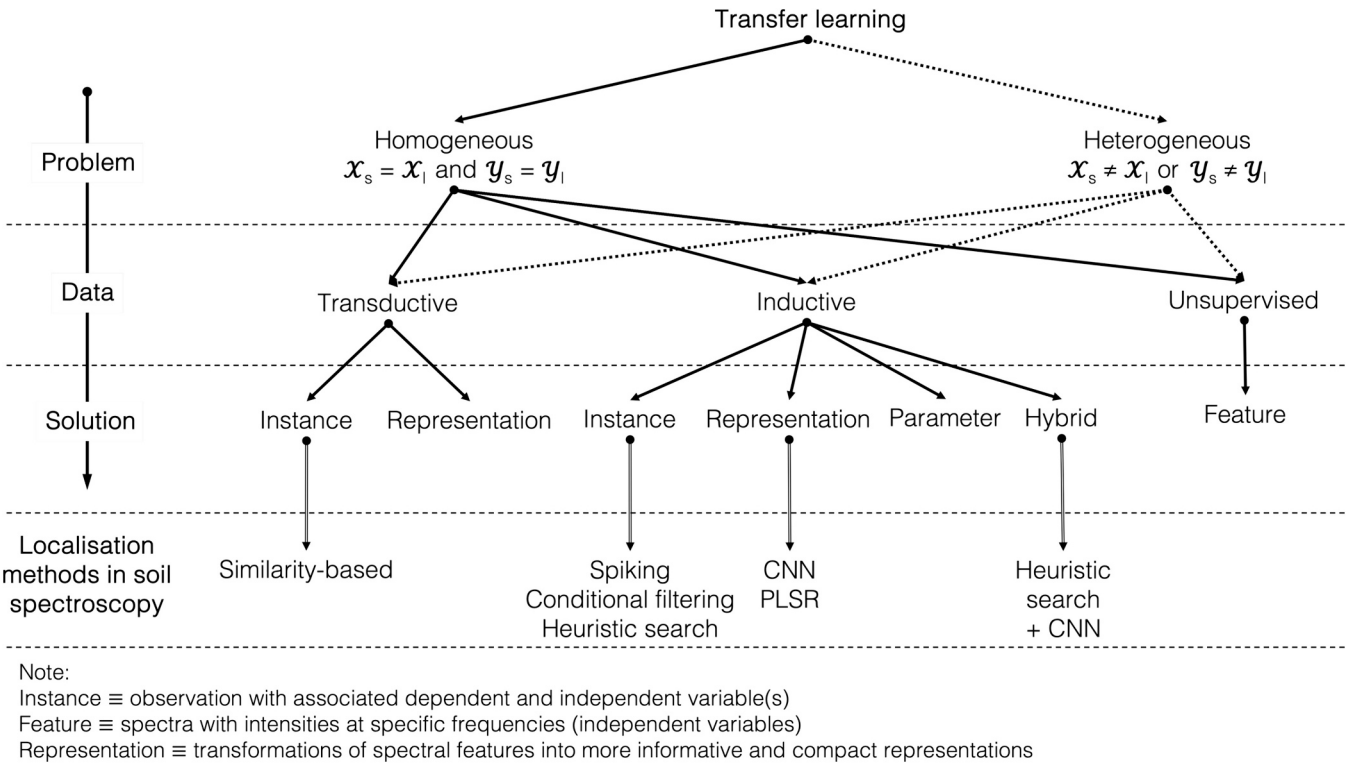

Fig. 2. Classification of transfer learning (TL). Dashed lines extending from Heterogeneous TL indicate that we have found no studies on heterogeneous TL in the soil spectroscopic literature.

3.2.2. The availability of local data for transfer learning

The availability of local observations with the measured soil property will determine whether the TL is transductive, inductive or unsupervised (Fig. 2). Transductive TL occurs when the measured response variable is only present in the source domain, and the TL can be on the observations (i.e., the instances) or the representations (i.e., the extracted spectral features that capture the underlying structure in the input spectra). The most widely known example of transductive TL is domain adaptation (e.g., Pan et al., 2011). Interested readers should see Niu et al. (2020) for a detailed description and possible algorithms and solutions. There are few examples of transductive TL in the soil spectroscopic literature (see Fig. 2 and Section 2).

Inductive TL occurs when the measured response variable is present in both source and local spectral domains. It aims to develop a predictive function with a small set of observations (with measured response variable) from the local domain to *induce* the local predictive function. Four approaches exist (Fig. 2), TL on instances, features, parameters, or combinations. Interested readers should see (e.g., Pan and Yang, 2010; Weiss et al., 2016; Niu et al., 2020) for detailed descriptions of inductive TL and the range of possible algorithms and solutions. There are examples of inductive TL in the soil spectroscopic literature (see Fig. 2 and Section 2).

Unsupervised TL refers to the case where neither the source nor the local spectral domains have data on the response variable (i.e., the soil property). There is limited application of unsupervised TL in soil spectroscopy. The only possible approach might be to aid with the clustering or dimensionality reduction of the local spectra with sufficient spectra from the source spectral library.

3.2.3. The possible transfer learnign solutions

Once the similarity between the source and local domains and data availability is known, the next decision will be the solution sought, which will depend on the specific undertaking. Experimentation and empirical evaluation are often needed to determine the most suitable approach.

Instance-based TL (Fig. 2) uses a small sample from the local domain (with or without measured response varaibles), to either re-weight the instances in the source domain or to extract relevant parts of the source data that can be reused with the small sample set to reduce the differences in the data distribution between the domains (Pan and Yang, 2010). One of the most popular methods uses the TrAdaBoost algorithm (Dai et al., 2007), which uses a few observations from the local domain to extract helpful information from the source domain by iteratively re-weighting the source observations. Instance-based methods can be useful when the measured soil property in the local domain is scarce or unavailable.

There are few examples of transductive instance-based TL in the soil spectroscopic literature; for example, the similarity-based localization methods (Fig. 1), may be classed as instance-based transductive TL (Fig. 2). Most of the existing spectroscopic localization methods (Section 2), can be categorized as inductive instance-based TL, e.g., spiking, conditional filtering, and heuristic search methods (Fig. 1). Spiking is an extreme case of instance-based TL, where all observations in the source spectral domain are reused to derive the local predictive function.

Representation-based TL (Fig. 2) assumes that the source and local spectral domains have representations in common, i.e. that the representations (of the spectral features) are domain invariant (Tzeng et al., 2014). It aims to transfer the 'learned' representations from a model that is typically trained on a large dataset to the local domain. Using the pre-trained representations, the local model can benefit from the information captured during pre-training. Representation-based TL may be a good choice when the pre-trained model captures general patterns and features that are relevant across the domains (Yosinski et al., 2014). Representation-based transfer can be effective when there is no measured soil property data in the local domain and the source and local domains have similar high-level features.

Weiss et al. (2016) describe asymmetric and symmetric feature TL. Asymmetric feature transformations are used when the conditional distributions of the source and local response variable are the same so that transformation can occur without context feature bias. Symmetric feature transformations help find underlying structures in the data by transforming both domains to a common predictive low dimensional latent feature space while reducing the marginal distribution between the domains (Weiss et al., 2016). Examples of feature transfer in the localization of spectroscopic soil modelling (Fig. 1 and Fig. 2) include the reusing of the representations in neural networks (Liu et al., 2018) and PLSR (Ng et al., 2022).

Parameter TL (Fig. 2) uses shared model parameters or prior distributions of hyperparameters from source and target domains. Some or all pre-trained parameters are transferred to the local model in this case. Parameter transfer may be useful when the source and local domains have similar low- and high-level features, and the pre-trained model's parameters can be directly applied to the local undertaking. This approach is typically used when there is sufficient local data for fine-tuning the pre-trained model's parameters. This transfer type is only suitable when the dissimilarity between the domains is small (Niu et al., 2020). There are no studies in soil spectroscopic modelling that use parameter transfer, however, parameter transfer has been used for the classification of remote sensing images (Ma et al., 2021).

Emerging approaches in the TL literature include hybrid methods that simultaneously transfer instances, features or shared parameters and relational TL methods that aim to 'learn' the common relationships between the source and target domains. They have yet to be explored in soil science. Shen et al. (2022) combined instance- with feature-transfer using the RS-LOCAL 2.0 algorithm and representation-transfer by fine-tuning CNNs. Further research on these methods is needed.

3.2.4. Positive, negative and zero transfer

The distinction between positive, negative and zero transfer is a point to note. TL depends on the relevance and compatibility of the information transferred from the source to the local domain. Therefore, positive TL occurs when the information gained from the source domain improves the performance of the predictive function in the local domain, $f_l(\cdot)$. In this case, the transfer enhances the 'learning' by $f_l(\cdot)$ and improves the model's generalization and the soil property estimation in the local domain. Conversely, negative TL occurs when the information from the source domain degrades the performance of $f_l(\cdot)$. Negative TL can occur when the information from the source domain is irrelevant or incompatible with the local domain (Wang et al., 2019). Of course, we should aim for positive TL and avoid negative TL because it can produce inaccurate or erroneous results, which can mislead decision-making. Although, there are studies in soil spectroscopic modelling literature that report on the variable and poor performance of localization (i.e. spectroscopic TL) (e.g. Guerrero et al., 2010; Seidel et al., 2019; Ng et al., 2022; Shen et al., 2022), we have not found studies that *explicitly* diagnose and address negative TL. Zero transfer refers to a situation where the information from the source does not offer substantial benefits or improvements in the performance at the local domain. It occurs either when there is significant dissimilarity between the domains or no helpful transferable information.

3.3. Implementing transfer learning

For the effective implementation of TL, one should answer three fundamental questions: *when to transfer*, *what to transfer*, and *how to transfer* (Pan and Yang, 2010). The order in which these questions are addressed can vary depending on the context and problem. Unless there are resource limitations, prior knowledge or expertise, it may be practical to sequentially address the *when*, *what*, and *how*, for a systematic and practical use of TL. (1) *When to transfer* emphasizes that one should not use TL in all situations. TL should only be used when it improves the accuracy of the local predictive function. There may be situations when

the transfer does not improve or even degrades the accuracy of the local estimates, resulting in zero transfer or negative transfer where the estimation fails. (2) *What to transfer* asks to identify the information that should be transferred from the source to the local domain. Whether transferring the information contained in the observations, representations, or both, the aim is to leverage the helpful information from the source domain to improve the estimation in the local domain and achieve positive TL. Therefore, knowing what to transfer allows us to use the soil property data, the spectra, and learned representations to improve the accuracy, efficiency and generalization of the local predictive function. (3) *How to transfer* pertains to the strategies and techniques for transferring the information effectively from the source to the local domain (Fig. 2). Selecting the most appropriate techniques will ensure optimal transfer, maximizing the use of the transferred information.

We have discussed spectroscopic localization and proposed that TL can elegantly describe the problem and help to develop robust and practical solutions systematically. Next, we report on experiments that test the implementation of a TL method with a large and diverse global vis–NIR SSL and data from 12 local sites worldwide. Following our framework above, the TL problem is homogeneous because the SSL (i.e. the source domain) and the local data (i.e. the target or local domain) share the same response variable and the same spectral range; it is inductive because several local samples analyzed for SOC help to induce the transfer; and it is instance-based because the method transfers helpful observations from the SSL to assist the local modelling at each site.

4. Data and methods

The global SSL used in this work (GSSL 2.0) encompasses a subset of the global soil spectral library (GSSL) described by Viscarra Rossel et al. (2016), the World Soil Information (ISRIC) spectral library (World Agroforestry (ICRAF) and International Soil Reference Information Centre (ISRIC), 2021; Shepherd and Walsh, 2002; Shepherd et al., 2003), the European Land Use and Coverage Area frame Survey (LUCAS) database (Stevens et al., 2013), the Mediterranean spectral database (i-BEC et al., 2019; Tziolas et al., 2019), the Rapid Carbon Assessment Program (RaCA) (Wills et al., 2014), and the Chinese spectral library (Shi et al., 2015). Thus far, it holds 52,742 spectra. Table 1 summarises the spectroscopic information of each database. Interested readers should see the relevant publications above for more specific details.

To combine the reflectance spectra from the different libraries with different spectral ranges, resolutions, and wavelength intervals (Table 1), we interpolated the spectra to a standard 10 nm wavelength interval using a local polynomial regression (Cleveland, 1981). The interpolation also improved the signal-to-noise ratio and reduced the dimensionality and redundancy in the spectra.

Fig. 3a shows the spatial distribution of the samples in the GSSL 2.0. The reflectance spectra of the GSSL 2.0 were transformed to apparent absorbance using $A = log_{10}R^{-1}$, where R is the reflectance, and then standardised using the standard normal variate (SNV) transformation (Barnes et al., 1989). Fig. 3b shows the pre-processed spectra by continent.

The SOC content in the different datasets was measured using different analytical methods. The SOC content of the LUCAS samples was measured by dry combustion (Orgiazzi et al., 2018), and those of the Chinese samples by H_2SO_4-dichromate oxidation (Shi et al., 2015). The SOC of the samples from the GSSL and ISRIC sets was measured using different methods, including Walkely-Black, oxidation with H_2O_2, loss on ignition, CHN pyrolysis, Tyurin method, Springer-Klee, and dry combustion (Viscarra Rossel et al., 2016). The SOC content of the Mediterranean samples was measured using Walkely-Black and oxidation with H_2O_2 and loss on ignition (Tziolas et al., 2019). The GSSL 2.0 holds data from diverse soils, including organic soils in temperate regions in the northern hemisphere (Fig. 3a). The SOC content of the soils in the GSSL 2.0 ranges from 0.01% to 58.68%.

4.1. The local data

We tested local datasets from 10 countries (Canada, USA, Brazil, Sweden, Spain, Israel, Nigeria, Zimbabwe, China, and Australia) and a region of Antarctica (Ross Dependency), independent from the GSSL 2.0, originating from each continent and from soil that is used for different purposes including cropping, grasslands, forests, and shrublands. These local data represent soil at within-field, field, farm, and regional scales.

The mean SOC of the local data ranges from 0.07% to 11.13% (Table 2). The Canadian site has the largest variation in SOC content (0.10% to 53.30%), whereas SOC in the Ross Dependency has the smallest (0.01% to 0.46%).

The local soil samples were measured with different instruments but

Table 1
Summary of spectrometers used to measure the samples in the GSSL 2.0.

Instruments	ISRIC	LUCAS	GSSL	Mediterranean	Chinese
Spectrometer	ASD FieldSpec Pro	XDS RCA	ASD spectrometers	[a]ASD FieldSpec [b]PSR + 3500	ASD FieldSpec Pro
Manufacturer	Malvern Panalytical	FOSS	Malvern Panalytical	[a]Malvern Panalytical [b]Spectral Evolution	Malvern Panalytical
Detectors (nm)	Silicon: 350–1000,	Silicon: 400–1100,	Silicon: 350–1000,	[a]Silicon: 350–1000 [b]Silicon: 350–1000	Silicon: 350–1000
	InGaAs: 1001–1800,	PbS: 1100–2500,	InGaAs: 1001–1800;	[a]InGaAs: 1001–1800; [b]Photodiode: 970–1910;	InGaAs: 1001–1800;
	InGaAs: 1801–2500		InGaAs: 1801–2500	[a]InGaAs: 1801–2500 [b]Photodiode: 1900–2500	InGaAs: 1801–2500
Resolution (nm)	3 at 700,	2	3 at 700,	[a]3 at 700 [b]3 at 700	3 at 350–1000
	3 at 700,	2	3 at 700,	[a]3 at 700 [b]2.8 at 700	3 at 350–1000
	10 at 1400, 2100		10 at 1400, 2100	[a]10 at 1400, 2100 [b]8 at 1500; 6 at 2100	10 at 1000–2500
Wavelength range (nm)	350–2500	400–2500	350–2500	[a]350–2500 [b]350–2500	350–2500
Sampling interval (nm)	1	0.5	1	[a]1 [b]1	1

Part of the soil samples in the Mediterranean library were measured with an ASD[a] spectrometer and part with a PSR + 3500[b] spectrometer.

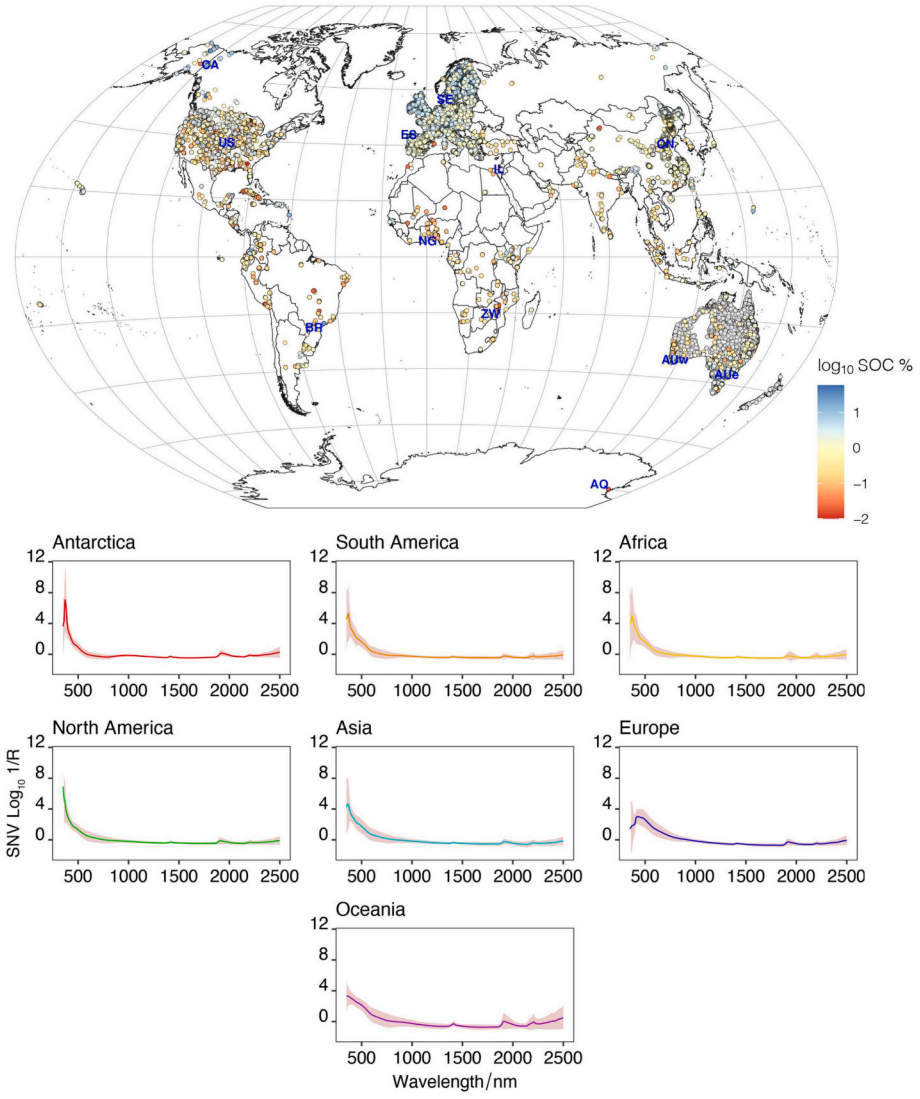

Fig. 3. (a) Spatial distribution of the GSSL 2.0, colored by total soil organic carbon (SOC%). Note that grey discs represent missing SOC data, but the GSSL 2.0 is likely to hold other soil properties data at these sites. (b) Average and standard deviation $log_{10}R^{-1}$ spectra by continent, pre-processed by the standard normal variate (SNV).

Table 2
Descriptive statistics of the total soil organic carbon (SOC) content in the GSSL 2.0 and in each of the 12 local sites. *K* represents the number of observations in the GSSL 2.0, *N* number of observations in the local data, S.d.is the standard deviation of the SOC contents, Min. the minimum value of SOC, $Q_{0.25}$ is the first quartile, $Q_{0.75}$ the third quartile, and Max. is the maximum SOC. Soil types are from the FAO-UNESCO soil type classification (Food and Agriculture Organization of the United Nations and Unesco, 2003): A = Acrisols, B=Cambisols, E = Rendzinas, F=Ferralsols, I = Lithosols, L = Luvisols, N=Nitosols, Q = Arenosols, R = Regosols, V=Vertisols, W=Planosols, X = Xerosols, Y=Yermosols.

Databases				*K*	Mean	S.d.	Min.	$Q_{0.25}$	Median	$Q_{0.75}$	Max.
GSSL 2.0				52,742	2.84	6.19	0.01	0.53	1.26	2.54	58.68
Local data	**Code**	**Area** $\left(km^2\right)$	**Soil type**	*N*	**Mean**	**S.d.**	**Min.**	$Q_{0.25}$	**Median**	$Q_{0.75}$	**Max.**
Antarctica	AQ	3	–	54	0.07	0.09	0.01	0.02	0.03	0.07	0.46
Brazil	BR	10	F	899	0.77	0.42	0.06	0.47	0.64	1.10	2.97
Canada	CA	8208	B	76	11.13	18.69	0.10	0.30	0.70	11.28	53.30
China	CN	165	A	135	1.75	0.58	0.89	1.41	1.59	2.05	4.54
E. Australia	AUe	61	W	100	0.64	0.79	0.05	0.19	0.32	0.75	3.78
Israel	IL	5558	E,L,R,X,Y	146	1.52	1.27	0.03	0.52	1.25	2.35	6.43
Nigeria	NG	12,248	I,L,N	142	0.40	0.48	0.02	0.13	0.29	0.49	3.84
Spain	ES	260	B	107	0.89	0.71	0.23	0.56	0.70	0.97	5.81
Sweden	SE	0.8	B,W	108	2.25	0.57	1.31	1.78	2.23	2.64	4.47
USA	US	0.1	W	216	1.62	0.63	0.78	1.11	1.43	2.07	3.74
W. Australia	AUw	75	W	108	0.56	0.78	0.03	0.10	0.23	0.53	3.25
Zimbabwe	ZW	15,521	I,L,N,Q	91	0.41	0.46	0.02	0.11	0.20	0.51	2.45
Median		113		108	0.83	0.60	0.05	0.38	0.67	1.04	3.81

same make of spectrometer (ASD spectrometers, Malvern Panalytical, Worcestershire, United Kingdom) and we preprocessed the spectra in the same manner as the GSSL 2.0 (see above). Fig. 3a shows the locations of each local site and Fig. 4 their spectra.

4.2. Experiments

We designed our experiments using the GSSL 2.0 and local datasets (described above). For each of the 12 sites, we compared estimates of SOC using local models with a different number of representative local data, n (local_n), and transfer set models with $n+k$ observations (transfer_{n+k}), where k represents a subset of the GSSL 2.0, which contains K observations. We also tested the effect of PLSR and ML algorithms on the modelling and the stability of the local_n and transfer_{n+k} models. We used estimates from global models with K observations and local models with all N local observations as benchmarks, assuming that the global models would result in the least accurate estimates, while the local models with all N data would produce the most accurate estimates. Below, we detail the procedures.

4.2.1. Selecting different $\mathbf{local}_n$ subsets

To test the effect of (small and affordable) sample size on local modelling, for each of the local datasets with N observations from the ten countries and the Ross Dependency, we selected n representative samples using the Kennard-Stone algorithm (Kennard and Stone, 1969). Thus, for each of the 12 sites, we produced 10 'local_n' subsets with $n = 5, 10, 15, \ldots, 50$ observations for modelling. The remaining $(N - n)$ data served as the independent set to validate the models (see Section 1).

4.2.2. Selecting $\mathbf{transfer}_{n+k}$ subsets

To test the value of using the GSSL 2.0 for local modelling, we used the different local_n data from the ten countries and the Ross Dependency to perform an RS-LOCAL 2.0 search (Lobsey et al., 2017; Shen et al., 2022) and transfer $k \approx 100$ instances from the GSSL 2.0. Detail on the algorithm can be found in Lobsey et al. (2017) and Shen et al. (2022). Briefly, RS-LOCAL 2.0 has two key components: (i) it uses repeated simple random resampling to search over the large spectral library and select instances that are useful for modelling locally, and (ii) the selection is based on the performance of PLSRs derived with the random subsets, which help to account for the covariation between the response variable and the spectra. Thus, the algorithm keeps only the k instances from the large spectral library that produce the most accurate local models when assessed on the local n observations. Once selected, the k are combined with the n data to construct the transfer_{n+k} sets. Our implementation of RS-LOCAL 2.0 sets its b parameter, the number of times a sample is drawn from the GSSL 2.0, on average, during the re-sampling, to $b = 80$, following recommendations in Lobsey et al. (2017) and Shen et al. (2022).

4.2.3. Modelling with different algorithms

To assess the effects of different algorithms on the modelling with the global, local_N, local_n, and transfer_{n+k}, we used PLSR (Wold et al., 2001), the regression tree method CUBIST (Quinlan, 1992), SVM with a radial basis function (Vapnik et al., 1996), and optimised one-dimensional CNN (Shen and Viscarra Rossel, 2021). Readers are directed to those publications for detail on the methods. Examples of their use in spectroscopic modelling can be found in Viscarra Rossel and Behrens (2010) and Shen and Viscarra Rossel (2021).

Implementation. Before modelling, we centred the spectra in the global, local_N, local_n, and transfer_{n+k} sets. Each regression method has several hyperparameters to be optimised. For each method, we set the optimisation objective to minimize the root mean squared error (RMSE) derived from 10-fold cross-validation of the global and local_N, and 5-fold cross-validation of the local_n, and transfer_{n+k} models. For PLSR, we optimised the number of PLS factors; for CUBIST, the number of committees and neighbours; for SVM, the cost (C) and sigma; and CNNs, the number of convolutions and fully-connected blocks and their internal hyperparameters. Optimisation of the CUBIST and SVM hyperparameters were performed using the differential evolution algorithm (Mullen et al., 2011) and optimisation of the one-dimensional CNNs was performed

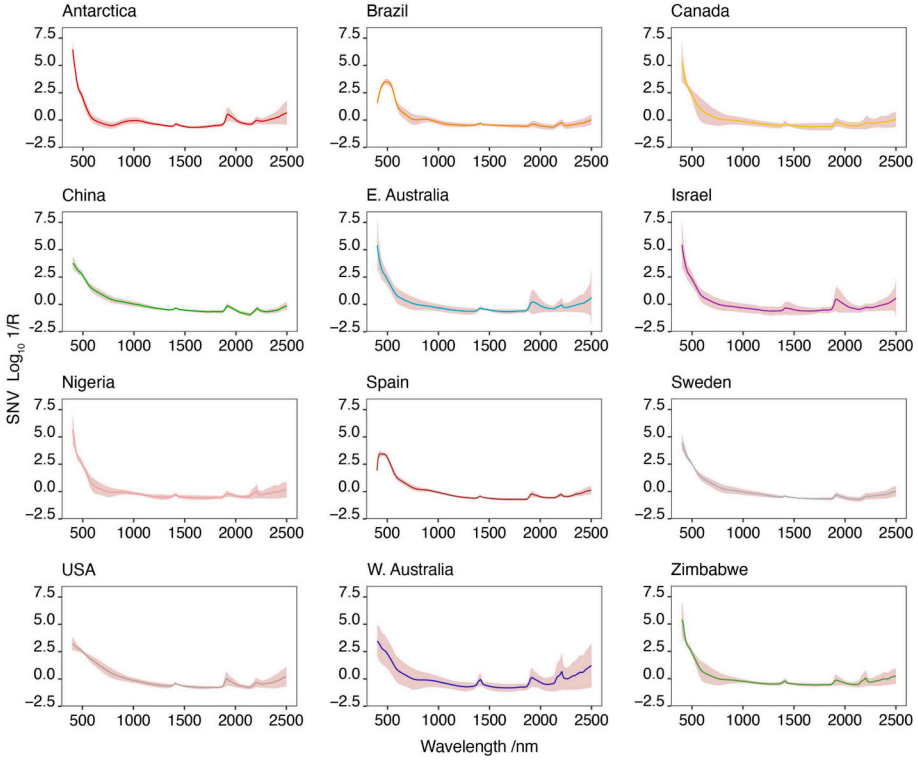

Fig. 4. Average and standard deviation $log_{10}R^{-1}$ spectra of each of the local sites, pre-processed by the standard normal variate (SNV).

using Bayesian optimisation with the Tree Parzen Estimators (Bergstra et al., 2011; Shen and Viscarra Rossel, 2021). We implemented PLSR, CUBIST, and SVM using the R software (R Core Team, 2022) and the caret library (Kuhn, 2008). The CNNs were developed in Python using the deep learning framework TensorFlow (Abadi et al., 2016).

Evaluating the accuracy of the models. We validated the global, local$_n$, and transfer$_{n+k}$ models by comparing their estimates of SOC to the measured $N - n$ observations from each of the ten countries and the Ross Dependency. The local$_N$ models were validated with a 10-fold cross-validation. We used the difference between the observed versus predicted SOC values to compute Lin's concordance correlation coefficient (ρ_c) (Lin, 1989), which helped to compare the different models, the RMSE to measure their inaccuracy, the mean error (ME) to measure their bias and the standard deviation of the error (SDE) to measure their imprecision. ρ_c is unit invariant and ranges from −1 to 1, making the direct comparison between sites possible. The RMSE, ME, and SDE explicitly characterize the models' estimation errors (Viscarra Rossel and McBratney, 1998).

4.2.4. *The stability of the* **local$_n$**, *and* **transfer$_{n+k}$** *models*

To test the stability of the SOC models that used the local$_n$ and transfer$_{n+k}$ data, we identified the algorithm that produced the best estimates of SOC and modelled the data with that method using 50 non-parametric bootstraps (Viscarra Rossel, 2007). The bootstrap uses samples drawn at random with replacement to assess the variations in the modelling that arise from data structurally similar to that under study, which could have plausibly arisen instead. We computed the mean and standard deviation of the 50 estimates to quantify the stability of the local$_n$ and transfer$_{n+k}$ models.

4.3. *Interpreting the transfer*

We interpreted the transfer from data, modelling and soil science perspectives to gain insights into the RS-LOCAL 2.0 transfer and the information carried from the GSSL 2.0 to the local sites. As stated above, TL aims to reduce the discrepancies in the marginal and conditional distributions of the data in the source and local domains. We used the principal component scores of the spectra as a proxy for the marginal distributions of the global, local$_n$, and transfer$_k$ data and to assess the RS-LOCAL 2.0 transfer. We also compared the conditional distributions of the SOC in the global, local$_n$, and transfer$_k$ datasets to assess the transfer with RS-LOCAL 2.0.

Because RS-LOCAL 2.0 uses PLSR to select the information to transfer, it considers the co-variation between the SOC and the spectra. Therefore, we were able to analyse the variable importance of the PLS models of the global, local$_n$, and transfer$_k$ data to assess if the information selected by RS-LOCAL 2.0 for the transfer helped derive the local predictive function. The transfer is positive if the spectra-SOC relationships of the local$_n$ and transfer$_k$ models are similar and related. If they are not, the local predictive function will be biased, no transfer will occur, and the transfer could be negative. We performed PLSRs on the global, local$_n$, and transfer$_k$ data and tuned the models using 5-fold cross-validation. Variable importance was calculated using the *varImp* function from the CARET library in the software R. The function calculates variable importance as a weighted sum of the absolute regression coefficients (Kuhn, 2008).

To determine if soil and environmental factors were responsible for, or at least contributed to the RS-LOCAL 2.0 transfer, we compared different attributes of the local$_n$ and transfer$_k$ samples. For the comparison, we plotted the coordinates (latitude and longitude) of the data to assess geographical similarities, then using a geographic information system (GIS), extracted values at the data locations from maps of: i) soil properties including bulk density, cation exchange capacity (CEC), water pH, and clay content, from the global soil grids (Poggio et al., 2021), ii) climate variables, mean annual temperature (MAT), and mean annual precipitation (MAP) (Fick and Hijmans, 2017), iii) soil types using the only currently available digitised global soil classification system (Food and Agriculture Organization of the United Nations and Unesco, 2003), and iv) a global land cover classification (Buchhorn et al., 2020).

5. Results

The GSSL 2.0 holds data from diverse soils, including organic soils in temperate regions in the northern hemisphere (Fig. 3a). In contrast, the local data originate mainly from agricultural fields, farms or areas with relatively small concentrations of SOC. The exception is the data from Canada, where the range in SOC is 0.1% to 53.3%.

The principal component analysis (PCA) scores of the pre-processed GSSL 2.0 spectra generally overlap, suggesting subtle differences in the mineral-organic composition of the soils from the seven continents (Fig. 5). The European spectra extend the feature space of the GSSL 2.0 (Figs. 5), with a more significant proportion of soils from temperate regions in the northern hemisphere (Figs. 3a). The projection of the local spectra from the 12 local sites onto the global (feature) space shows that the local spectra mostly fall within the space of the GSSL 2.0 (Fig. 5), indicating that the GSSL 2.0 contains spectrally similar samples for the local sites.

5.1. *Modelling with different algorithms*

The different algorithms used to model SOC content in the local$_n$ and transfer$_{n+k}$ data had only minor effects on the estimates. However, PLSR produced the most accurate estimates when the sample size was small ($n \leq 20$), and CUBIST was most accurate with larger sample sizes ($25 \leq n \leq 50$) (Fig. 6a).

Modelling with all of the global data produced inaccurate estimates ($\rho_c \approx 0$), regardless of the algorithm used (Fig. 6a). As might be expected, cross-validation with the local$_N$ data produced the most accurate estimates (mean $\rho_c = 0.81 \pm 0.14$ s.d., depending on the algorithm). Modelling with the transfer$_{n+k}$ data improved the results compared to using only the local$_n$ data, particularly when the models used a smaller sample size ($n \leq 35$) (Fig. 6a). Improvement started to level off around $n = 30$. There was little benefit of using TL for $n \geq 40$.

5.2. *Modelling with the* **local$_n$** *and* **transfer$_{n+k}$** *data*

For a more detailed comparison of the models with the local$_n$ and transfer$_{n+k}$ data, we compared the estimates of SOC in the 12 sites using only CUBIST (Fig. 6b, Table 3). On average, modelling with the transfer$_{n+k}$ data produced more accurate SOC estimates than using only the local$_n$ data. The differences in ρ_c are slightly larger at smaller sample sizes, but on the whole, the CUBIST estimates with the transfer$_{n+k}$ data were, on average, 13.3% more accurate (Fig. 6). Since the advantage of TL started to diminish at $n = 30$, we show the assessment statistics for CUBIST using $n = 30$ in Table 3.

The accuracy of the SOC estimates varied at the different local sites. The CUBIST models from Brazil, Nigeria, Israel, and Spain produced somewhat inaccurate estimates ($\rho_c < 0.65$), although those using the transfer$_{n+k}$ data were better (Fig. 6b, Table 3). Conversely, estimates for Canada, Sweden, the USA, and China were accurate ($\rho_c \geq 0.8$) and more so when the transfer$_{n+k}$ data were used. The transfer$_{n+k}$ data also improved the accuracy of the SOC estimates in Antarctica and eastern Australia. Overall, except for Israel and Western Australia, which exhibited negative TL, the SOC estimates with the transfer$_{n+k}$ were more accurate than those with only the local$_n$ data (RMSE, Table 3).

5.3. *Instability of the* **local$_n$** *and* **transfer$_{n+k}$** *models*

Models built with the transfer$_{n+k}$ data were more stable than the local$_n$ models (Fig. 6c). The instability of the local$_n$ models decreased somewhat as the number of samples n increased, as shown by the narrowing width of the standard deviations of their mean estimates. However, the transfer$_{n+k}$ models were consistently more stable

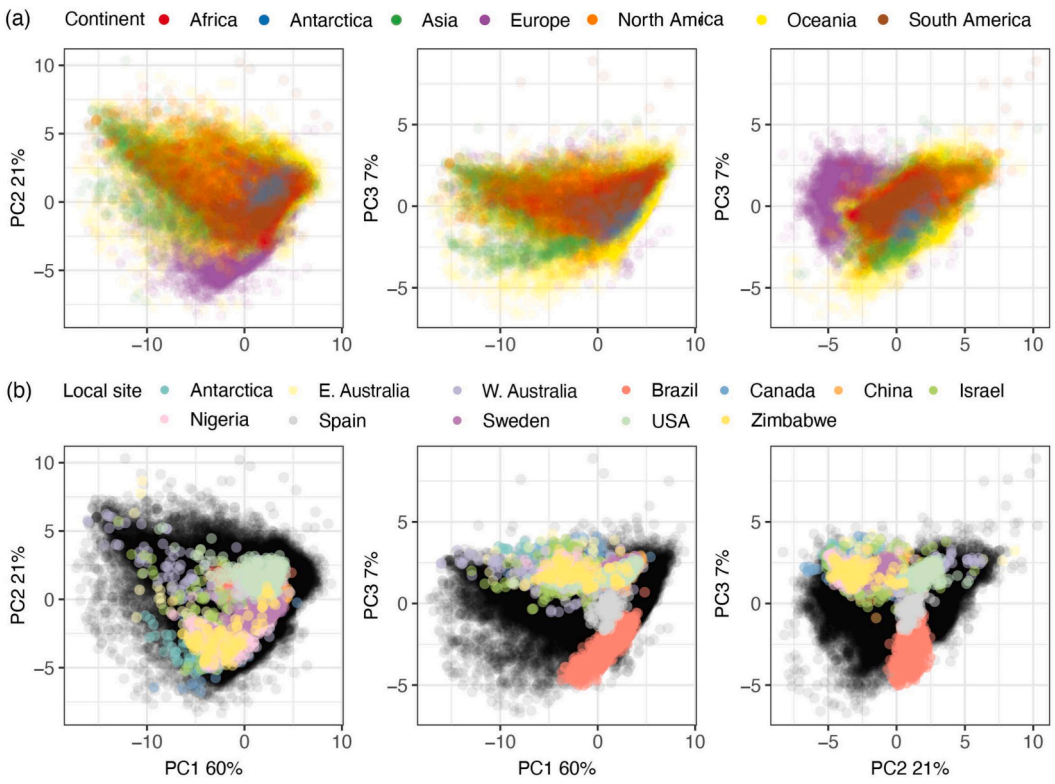

Fig. 5. Principal component analysis (PCA) of the data in the GSSL 2.0 colored by (a) continent and (b) local (local$_N$) data. In (b) the GSSL 2.0 spectra are shown in black.

regardless of sample size, as shown by the similar width of their standard deviations (Fig. 6c).

5.4. Interpreting the transfer: data and models

The statistical distributions of the SOC data in the GSSL 2.0 and the 12 local sites are different (Fig. 7a). However, the SOC distributions of the local$_n$ and transfer$_k$ data are more similar because the transfer with RS-LOCAL 2.0 helped to reduce the dissimilarity between the conditional distribution of the GSSL 2.0 and the local data.

The projection of the local$_n$ and transfer$_k$ data onto the spectral space of the GSSL 2.0 shows that RS-LOCAL 2.0 selects instances that enhance and extend the spectral space occupied by the local$_n$ observations (Fig. 7b). Instance-based TL with RS-LOCAL 2.0 produced similar spectral distributions in PCA space, implying a reduction of the dissimilarities between the marginal distributions of the GSSL 2.0 and the local spectra. A few Western Australian observations fell outside the GSSL 2.0 spectral space (Fig. 7b), indicating that the library does not hold observations similar to those from this site, causing the marginal distributions in the local$_n$ and the transfer$_k$ data to be different, resulting in negative TL (W. Australia, Table 3).

Generally, for each of the 12 sites, the variable importance of the local$_n$ PLSRs were more similar to those from the transfer$_k$ PLSRs than the variable importance of the global model (Fig. 8). TL with RS-LOCAL 2.0 selected instances from the GSSL 2.0 that shared similar spectra-response relationships with the local$_n$ data. The magnitude of the variable importance between the local$_n$ and transfer$_k$ data from Western Australia are very different, though the models used similar wavelengths. The variable importance of the Nigerian local$_n$ and transfer$_k$ models were the most different. Differences in the magnitude of the importance or the wavelength regions indicate differences in the spectra-response relationships of the local$_n$ and transfer$_k$ models, which contributed to their reduced performance and negative TL (Table 3).

5.5. Interpreting the transfer: geography and soil science

Exploring the geographical distribution of the transfer$_k$ data (Fig. 9) shows that geography alone does not explain the transfer by RS-LOCAL 2.0 from the GSSL to local sites. However, we can discern general patterns. The transfer$_k$ selection for each site comes from various locations worldwide, but mainly from the US, Europe, China, and Australia, which constitute the largest contributions of the data in the GSSL 2.0 (Fig. 3). The transfer$_k$ instances selected by RS-LOCAL 2.0 from the GSSL 2.0 extended beyond the geographic space of local$_n$ data, implying that observations from a location anywhere in the world are related to those from other places (presumably because they occur under similar pedo-climatic conditions) and benefit TL and the local estimation.

Unsurprisingly, the selection shows to be related to the SOC content of the soil at the sites. Australian samples are prominent in the transfer$_k$ samples of all 12 local sites (Fig. 9). Samples from the Midwest of the United States also tend to be more prominent in the selections, while there is no discernible pattern in the selection of samples from South America, Africa, and Asia (except China), possibly due to the sparsity of data on these continents (Fig. 3).

The selection of European samples tends to differ for the 12 sites. For example, the spatial distribution of the selected samples for Canada, Sweden, the United States, China, and Spain is similar, with a large selection of samples from all over Europe (Fig. 9). The average SOC of the soil at these sites ranges from 0.9 to 11.1% (Table 2). The selection of European samples is more sparse for the Brazilian and eastern Australian sites. The average SOC content of the soil at these sites ranges from 0.64 to 0.77% (Table 2). The selection for Nigeria and Zimbabwe are primarily from southern Europe and the Mediterranean (Fig. 9), and SOC content of the soil at these sites ranges from 0.4 to 0.56% (Table 2). The selection for the Antarctica site has the least number of European samples, and the SOC content of the samples is 0.1%.

In Fig. 10, we show the soil type of the samples from each local site and the soil type of the transfer$_k$ data. The soil types of the transfer$_k$

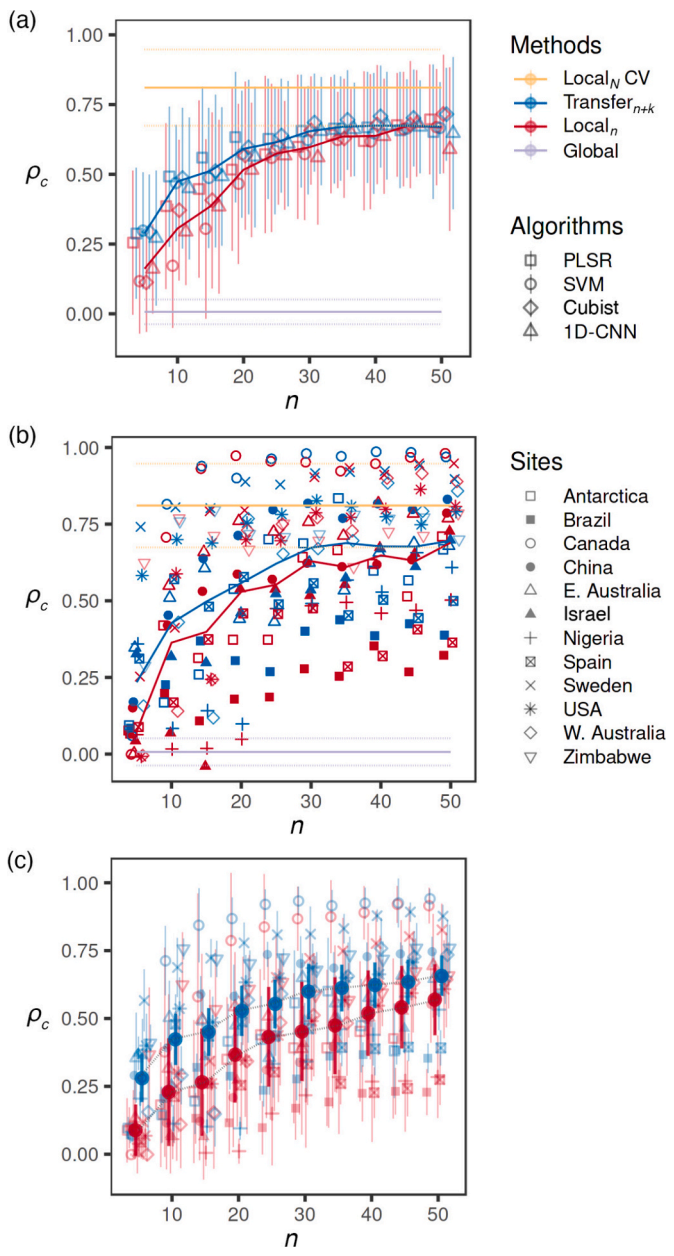

Fig. 6. Modelling assessment. (a) Concordance correlation, ρ_c, showing the estimates of SOC with the different algorithms and the global (K), local$_N$, local$_n$, and transfer$_{n+k}$ data sets. (b) Concordance correlation of the CUBIST estimates of SOC for each local site using the local$_n$ and transfer$_{n+k}$ data sets; (c) Bootstrap estimates of modelling stability with the local$_n$ and transfer$_{n+k}$ data sets using CUBIST. The error bars represent the standard deviation of ρ_c and the horizontal dashed lines in (a) and (b) the standard deviation of ρ_c for the estimates with the global (K) and the cross validated local$_N$ models.

selection did not all match the local soil types (Fig. 10). There were some similarities, however; for example, the soil types of the selected transfer$_k$ observations were similar to the soil types from the local sites in Nigeria, the USA, Western Australia, Zimbabwe, and to some extent Sweden and Israel (Fig. 10).

Fig. 11 shows the land cover at each local site and the land cover of the transfer$_k$ data. Generally, the most prominent land cover types of the selected transfer$_k$ observations matched the land cover at the local sites.

Climate, represented by MAT and MAP, did not directly affect the transfer (Fig. 12). The local$_n$ and transfer$_k$ data at four sites (eastern Australia, Spain, Sweden, and Zimbabwe) had similar MAP. The local$_n$ and transfer$_k$ data at only one site, Zimbabwe, had similar MAT. However, the MAP and MAT representing the global data were also similar to the transfer$_k$ data.

Soil properties did not seem to affect the selection of transfer$_k$ observations. In most cases, the soil property distributions of the local$_n$ and transfer$_k$ data are similar only if the distribution in the global data is also similar, e.g. bulk density (BD) in Nigeria and Spain, CEC in China, eastern Australia, and Spain, pH$_W$ in Nigeria, Sweden, and Zimbabwe. The clay content distribution of the local$_n$ and transfer$_k$ data was similar but different to the global data in only three sites (Nigeria, Spain, and Zimbabwe). Generally, however, climate and soil properties appear not to have affected the RS-LOCAL 2.0 selection (Fig. 12). However, the climate data are from global maps with coarse pixel resolutions, and the soil data are from extrapolated coarse resolution and uncertain soil property maps (see Methods).

6. Discussions

Despite the many regional, national, continental, and global SSLs developed and the remarkable evolution of ML and AI, they have yet to significantly impact the application and deployment of soil spectroscopy beyond research (Viscarra Rossel et al., 2022). Reasons for this might be the significant investment needed to develop SSLs with a comprehensive and precisely analyzed set of soil properties, and ML methods' need for large volumes of data for training the models. Obtaining soil analytical data on chemical, physical and biological properties is expensive (Viscarra Rossel and Bouma, 2016).

We now also understand that models developed using large 'global' vis–NIR SSLs cannot generalize to any local situation, so direct soil property estimation based on global models (e.g. Shepherd et al., 2022) will not generalize well or simply not work—see Table 3. At the core of this problem are the discrepancies between the marginal or conditional distributions of the data in the SSL and the local site. Like other predictive methods, spectroscopic modelling assumes that the data distributions in the training and prediction sets are similar. If the assumption is satisfied, the modelling should succeed. TL provides a framework and a continuously improving and rapidly developing set of methodologies that can help explicitly address the discrepancies in the data distributions.

We have shown that instance-based TL with RS-LOCAL 2.0 can help overcome the mentioned drawbacks by transferring relevant and beneficial information from a global spectral library, the GSSL 2.0, to improve the local modelling of SOC worldwide. TL with RS-LOCAL 2.0 improved the local modelling of SOC at 10 of the 12 local sites (Table 3). Negative TL occurred at two sites. Poor prediction at the Western Australian site is most likely because the GSSL 2.0 does not contain sufficient helpful data to estimate at this site (Fig. 7). For Israel, the spectra-SOC relationships in the local$_n$ and transfer$_k$ data sets differed, implying the transfer of observations that did not effectively improve the modelling, possibly because the SOC values in the local$_n$ data were imprecise.

When the TL was positive, RS-LOCAL 2.0 reduced the dissimilarity between the marginal and conditional distributions of the GSSL 2.0 and the local data. It resulted in spectroscopic models that were at least as accurate or better than entirely local models derived using the same number of local observations (Fig. 6b, Table 3). Local models with fewer than 30 observations produced estimates that were, on average, less accurate than those derived with an equivalent number of observations used in the transfer with RS-LOCAL 2.0. If one can only measure a small number of local samples, then RS-LOCAL 2.0 can help improve the cost effectiveness of soil spectroscopy. Therefore, instance-based TL with RS-LOCAL 2.0 and the GSSL 2.0 can facilitate the measurement and monitoring of SOC by minimizing the need for analytical measurements and reducing the cost of soil analysis.

RS-LOCAL 2.0 is an inductive instance-based TL method because it aims to derive a local predictive function with a small set of local observations. The selection of the transfer set is based on the performance of the

Table 3
Evaluation statistics for CUBIST models using the entire GSSL 2.0 (global, $K = 52{,}742$), all of the local data with N observations (see Table 2) and cross-validation (local_N), with $n = 30$ and $k = 100$.

Statistic	Model	Antarctica	Brazil	Canada	China	E. Australia	Israel	Nigeria	Spain	Sweden	USA	W. Australia	Zimbabwe
ρ_c	Global	0.02	0.00	0.05	0.11	−0.06	−0.01	0.02	−0.04	−0.04	0.00	0.01	0.00
	Local_N CV	0.53	0.64	0.97	0.85	0.79	0.78	0.71	0.81	0.96	0.93	0.94	0.80
	Local_n	0.64	0.28	0.95	0.62	0.62	0.52	0.48	0.48	0.90	0.79	0.77	0.67
	Transfer_{n+k}	0.69	0.40	0.98	0.82	0.76	0.54	0.49	0.56	0.92	0.83	0.67	0.72
RMSE	Global	1.04	1.15	20.16	1.00	2.15	2.38	1.70	0.96	1.43	0.98	6.67	3.08
	Local_N CV	0.08	0.31	4.20	0.29	0.49	0.76	0.35	0.40	0.15	0.22	0.26	0.29
	Local_n	0.07	0.40	5.70	0.43	0.66	0.97	0.44	0.46	0.24	0.38	0.51	0.30
	Transfer_{n+k}	0.06	0.37	3.68	0.29	0.61	0.99	0.40	0.42	0.23	0.32	0.56	0.29
ME	Global	0.86	0.56	−8.41	−0.51	1.08	0.57	1.30	0.31	−1.19	−0.29	4.34	1.89
	Local_N CV	−0.00	0.00	−0.02	−0.01	−0.02	0.01	0.02	−0.01	0.00	0.00	0.01	0.02
	Local_n	0.00	0.02	0.86	0.06	0.45	0.03	0.00	0.01	−0.10	−0.07	−0.06	−0.07
	Transfer_{n+k}	0.02	−0.06	0.39	−0.02	0.36	−0.15	0.03	0.05	−0.08	−0.05	−0.12	−0.05
SDE	Global	0.59	1.00	18.32	0.86	1.86	2.31	1.11	0.901	0.80	0.94	5.07	2.42
	Local_N CV	0.08	0.31	4.20	0.29	0.49	0.76	0.35	0.40	0.15	0.22	0.26	0.29
	Local_n	0.07	0.40	5.64	0.42	0.49	0.97	0.44	0.46	0.22	0.37	0.51	0.29
	Transfer_{n+k}	0.06	0.36	3.65	0.29	0.49	0.98	0.39	0.42	0.22	0.32	0.55	0.29

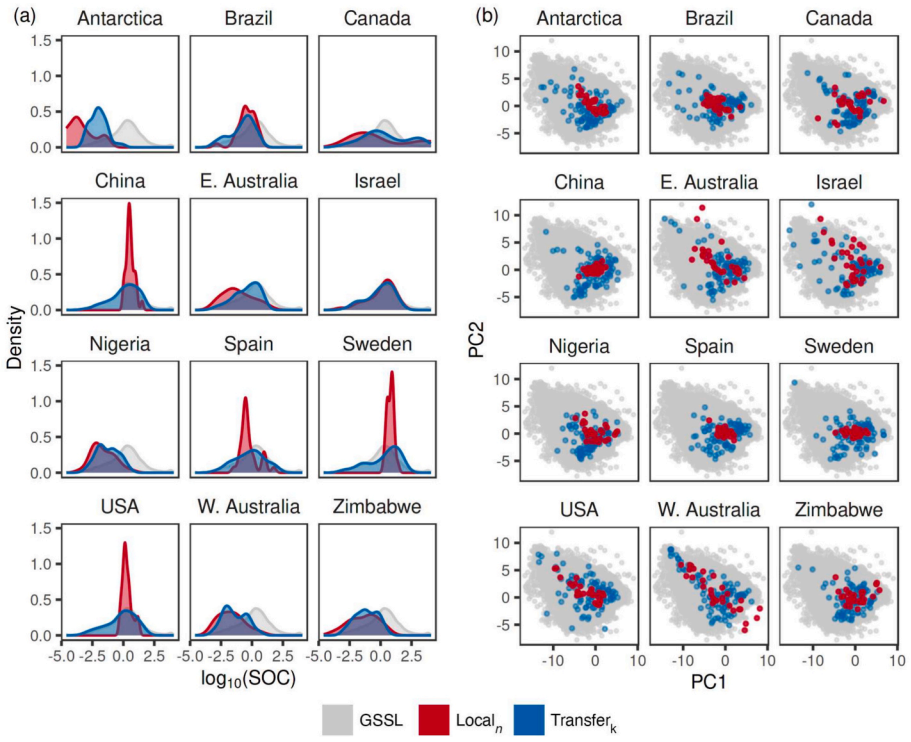

Fig. 7. (a) Density plots showing the distribution of soil organic carbon (SOC) in the GSSL 2.0 ($K = 52{,}742$), local_n ($n = 30$), and transfer_k ($k = 100$). (b) Scatter plot of the first two scores from a principal component analysis (PCA) of the global, local_n, and transfer_k spectra. The local_n and transfer_k spectra were projected on the GSSL's principal component space.

PLSRs that help account for the covariation between the SOC and the spectra, and the selected transfer_k instances from the GSSL 2.0, which share similar spectra-SOC relationships with the local observations (Fig. 8). Leveraging the soil spectra-SOC relationships helped to 'filter' the instances from the GSSL 2.0 to extract only the most relevant data for local modelling.

The GSSL 2.0 consists of spectra recorded for different projects, with different spectrometers (Table 1), and with SOC measurements made using different analytical methods. These inconsistencies in the spectra and the SOC analysis contribute to the differences in the marginal and conditional distributions of the data sets. However, RS-LOCAL 2.0 can reduce the differences in these distributions, reducing the effect of measurement inconsistencies. Therefore, the method should remove the need for separate 'calibration transfer' (Andrew and Fearn, 2004; Pittaki-Chrysodonta et al., 2021).

At each local site, RS-LOCAL 2.0 extracted instances from the GSSL 2.0 from places across the world that generally have similar mineral and organic matter composition (see Fig. 8), presumably because of similar pedoclimatic and management contexts. This aspect of the method is encouraging because samples collected and measured at one location can potentially help with modelling in other locations. Therefore, for spectroscopy to be a truly global, practical and cost-effective method, we should expand the GSSL 2.0 further and collect soil samples from under-represented areas to represent the vast diversity of global soils. A more extensive and diverse GSSL will provide richer and more helpful information for spectral TL.

We also carried out experiments to gain insight into the critical components of the transfer with RS-LOCAL 2.0 and to help users

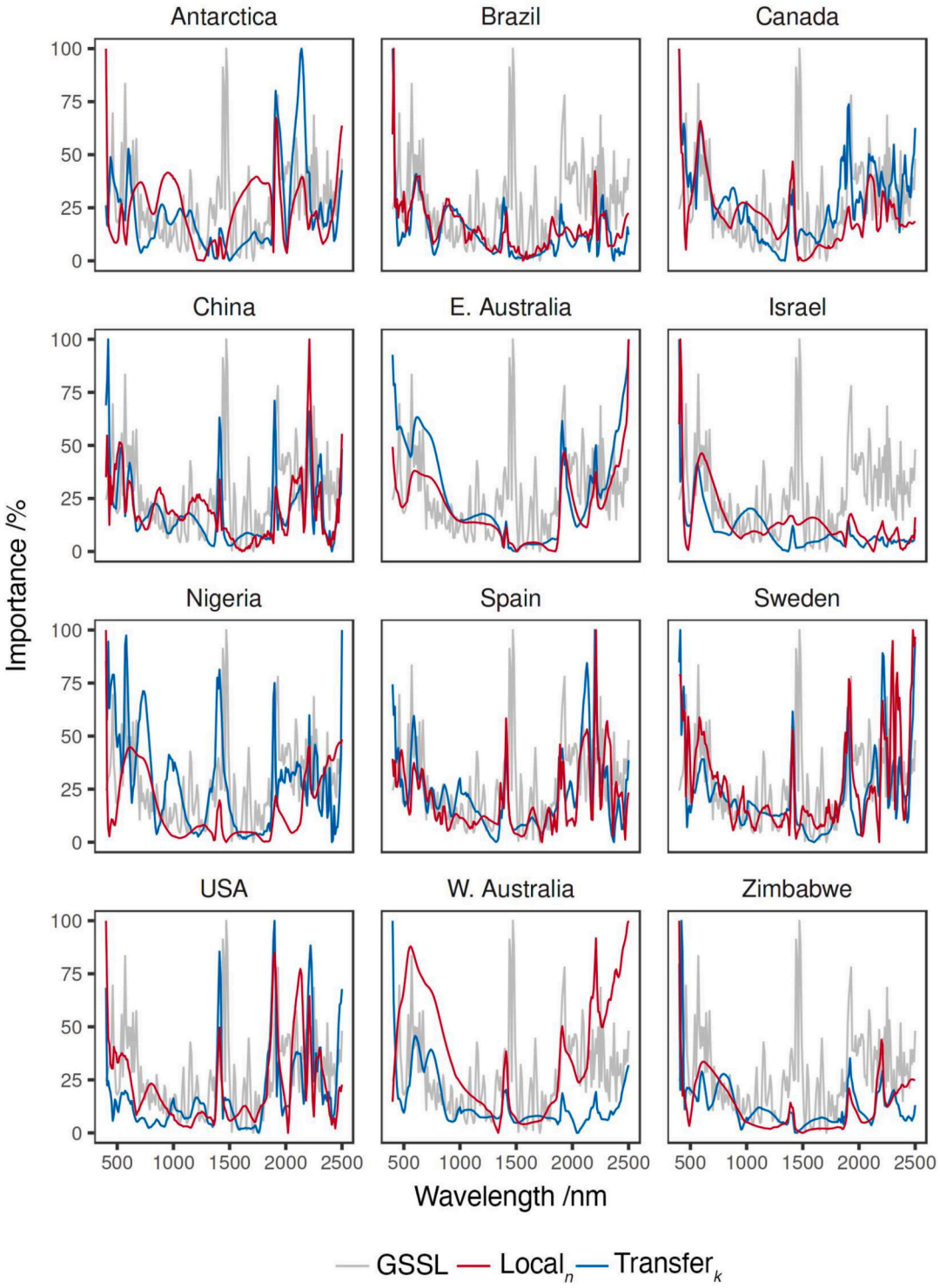

Fig. 8. Variable importance derived from modelling with the global (GSSL 2.0, $K = 52,742$), local$_n$ ($n = 30$), and transfer$_k$ ($k \approx 100$) data.

understand the algorithm. Developing and applying new interpretable and transparent methods is essential to encourage innovation and the adoption and development of soil spectroscopy. TL with RS-LOCAL 2.0 helped to diminish the dissimilarities in the marginal and conditional distributions between the GSSL 2.0 and the local data (Figs. 7). The selection of the instances from the GSSL 2.0 was relatively unconstrained by the geography of the local data (Fig. 9), climate, soil types (Fig. 10), or even soil attributes (Fig. 12) other than SOC. However, we understand the mismatch in the datasets' resolution and scale. The climate data are from global maps with 1 km pixel resolutions (Fick and Hijmans, 2017), the FAO soil map is at 1:5000000 scale and the classification is based on data from soil profiles (Food and Agriculture Organization of the United Nations and Unesco, 2003), and the soil property data are from extrapolated coarse resolution and uncertain soil property maps (Poggio et al., 2021). The selection of the transfer data appeared somewhat more affected by the land cover (Fig. 11), possibly because of the data's finer, 100 m pixel resolution (Buchhorn et al., 2020), and because of the direct effect of land cover on SOC concentrations.

Our results support the development of large and diverse SSLs, not to

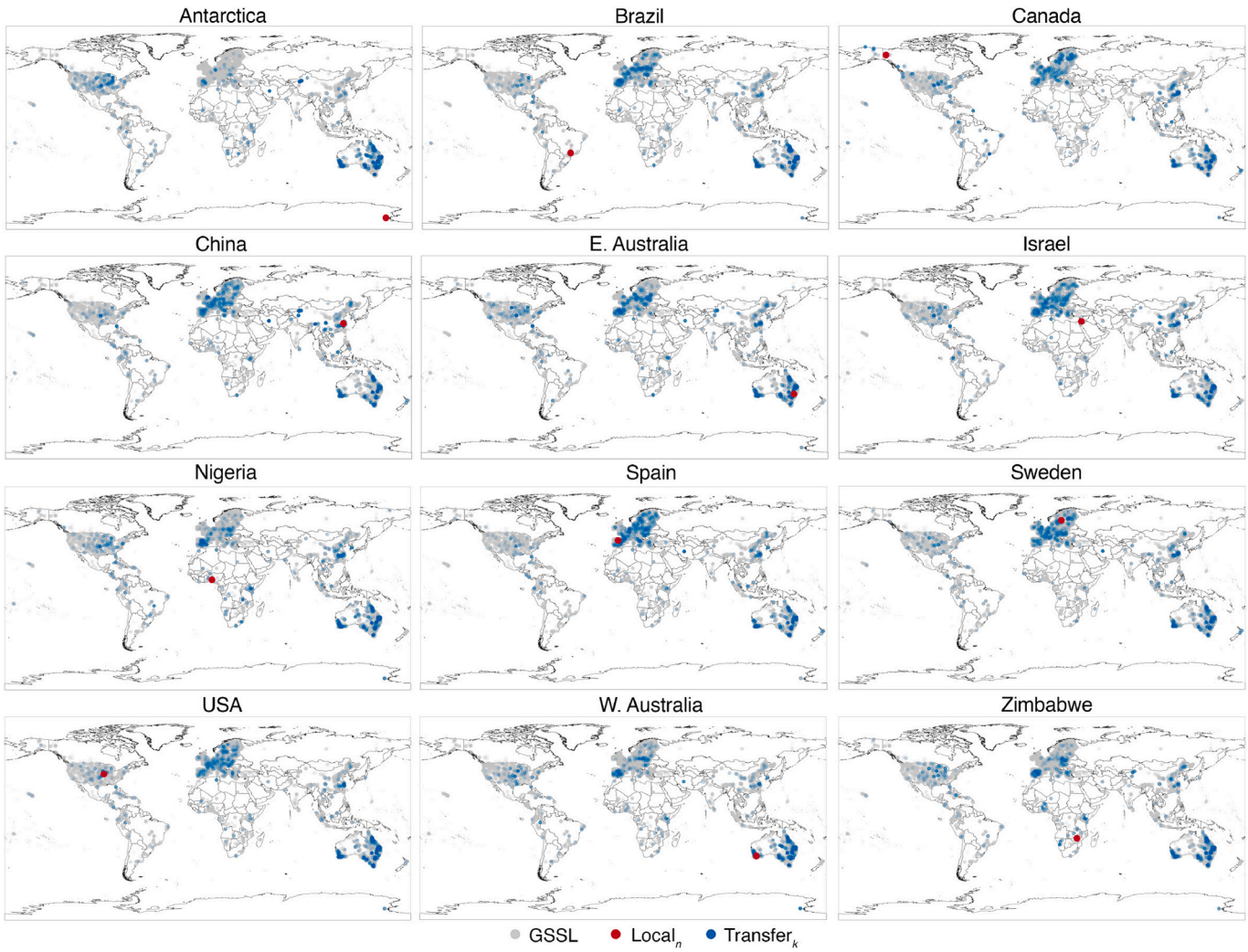

Fig. 9. Geographic locations of the local$_n$ (red dots) and transfer$_k$ (blue dots) observations.

derive global predictive functions but as a source of information for spectroscopic TL. We postulate that the larger and more diverse the global spectral library, the more likely it is to hold helpful information that can be transferred to derive accurate local soil spectroscopic models anywhere in the world. Therefore, maintaining and expanding global SSLs like the GSSL 2.0 to better represent the vast diversity of soils worldwide is essential and valuable for developing soil spectroscopy and cost-effective soil assessments and monitoring.

7. Future directions

The development and application of TL are necessary for the adoption of soil spectroscopy and its practical and cost-effective application for soil assessments and monitoring. We have shown that TL can help address the localization problem in soil spectroscopy. There is potential for developing innovative new methods for spectroscopic TL under the proposed framework (Fig. 2). Below, we briefly propose directions for the development of future research:

- Computationally efficient TL methods (Fig. 2) for better addressing the disparity in the marginal and conditional distributions and better capturing the spectra-response relationship of the source and local domain data.
- Inductive TL methods that can dynamically adapt to the local domain's size, diversity or even changing conditions. For instance, methods that use incremental learning and continuously update the model with new local data could be helpful in continuous soil sensing and mapping applications (e.g. mobile sensing platforms).
- Methods for adapting models across domains using domain adaptation algorithms, including those that combine domain adaptation with deep learning (e.g., deep adaptive neural networks, or DANN) (Tzeng et al., 2014). These techniques can help align the spectral distributions between the domains, considering the variations caused by, e.g., instrument calibration, measurement conditions, and environmental differences.
- Multi-task learning in spectroscopic TL. By jointly training the model on related soil properties (e.g. compositional data such as clay, sand and silt contents, or the organic carbon fractions), the model can leverage shared representations and transfer information across the data to improve local estimation.
- Fusing spectra from multiple sources or modalities. For example, combining spectra from different sensors (e.g. vis–NIR with laser induced breakdown spectroscopic (LIBS) spectra), or with auxiliary data, such as other soil properties, satellite imagery, climate and other environmental data (Yang et al., 2019, 2022), could provide complementary information for TL. Developing effective spectral fusion techniques can enhance the model's ability to capture diverse information from multiple data.
- Unsupervised learning techniques such as self-supervised learning (Zhai et al., 2019) or co-training (Ning et al., 2021) for leveraging the

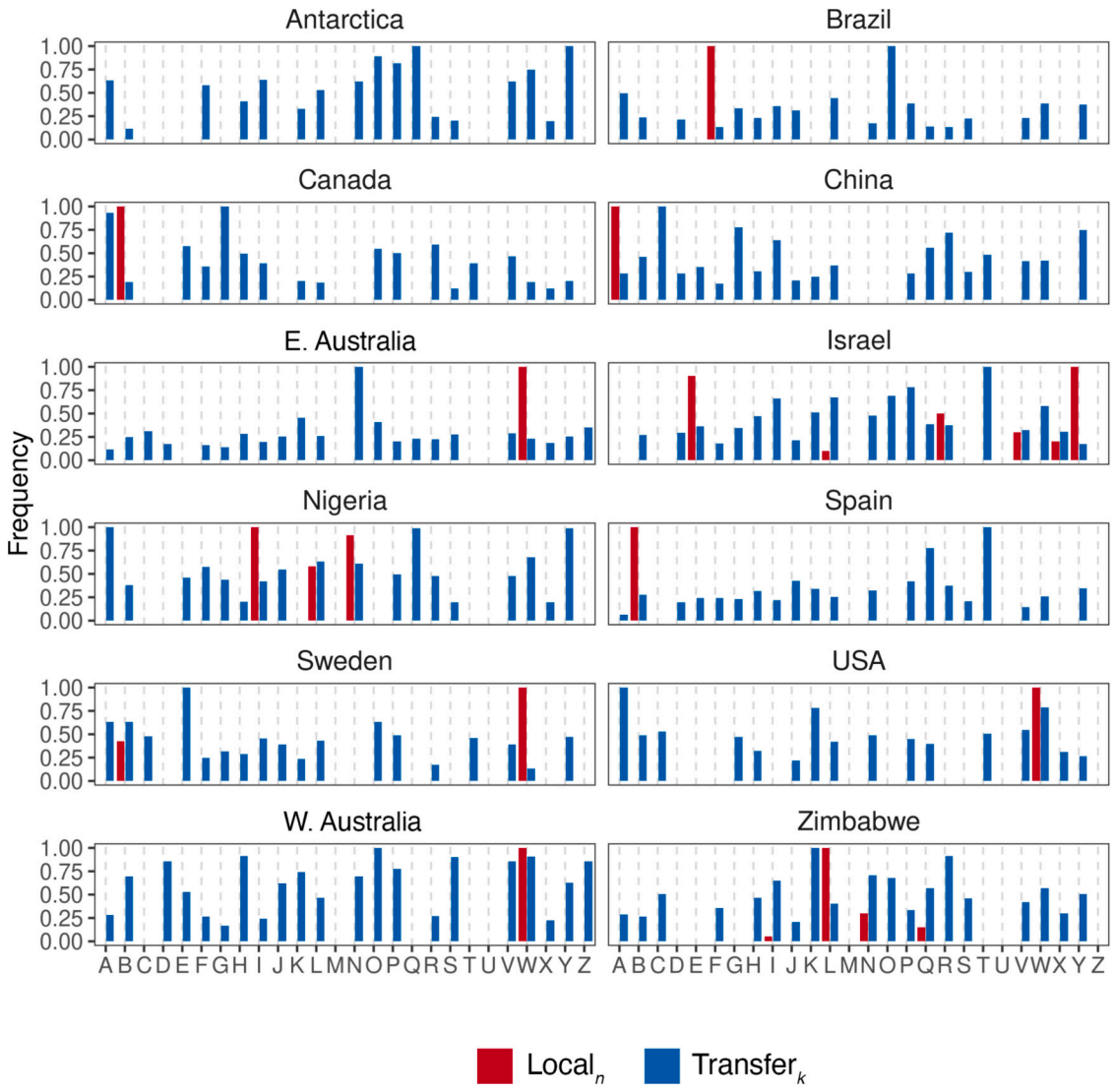

Fig. 10. The soil type(s) of the local$_n$ ($n = 30$) and transfer$_k$ ($k \approx 100$) data. Soil types are from the FAO-UNESCO soil type classification (Food and Agriculture Organization of the United Nations and Unesco, 2003): A = Acrisols, B=Cambisols, C=Chernozems, D=Podzoluvisols, E = Rendzinas, F=Ferralsols, G = Gleysols, H=Phaeozems, I = Lithosols, J = Fluvisols, K=Kastanozems, L = Luvisols, M = Greyzems, N=Nitosols, O=Histosols, P=Podzols, Q = Arenosols, R = Regosols, S=Solonetz, T = Andosols, U = Rankers, V=Vertisols, W=Planosols, X = Xerosols, Y=Yermosols, Z = Solonchaks. Soil type frequencies for the local$_n$ were derived from sample counts. Because the sample counts of the soil types in the GSSL 2.0 are imbalanced, counts of the soil types in the transfer$_k$ were normalised by the total of each soil type in the GSSL 2.0. The largest counts in the local$_n$ and transfer$_k$ data were scaled to 1.0.

abundance of 'unlabelled' local spectra (i.e. local spectra without measured soil properties) to improve TL.

- Extend the TL methods and those above to hyperspectral remote sensing, which involves a high-dimensional 2-D local spectral feature space. Similarly, extend spectroscopic TL to and its implementation with digital soil mapping.
- Methods to identify when TL will be positive, zero or negative. Positive TL is what we aim for. The methods can help determine whether TL is beneficial.
- Methods for interpretable TL to understand what information is reused in the transfer and how it helps the prediction.
- Easy-to-use software tools that enable the implementation of TL for globally-distributed, practical and cost-effective soil spectroscopy.

8. Conclusions

We reviewed current methods for localizing spectroscopic modelling and argued that localization is a typical TL problem. Then we reviewed and defined TL in the context of soil spectroscopic modelling and proposed that the proposed framework can guide the development of new methods in soil spectroscopy. We applied an instance-based TL method for soil spectroscopy with the RS-LOCAL 2.0 algorithm and used the GSSL 2.0 and local spectra from sites within ten countries worldwide and The Ross Dependency. Our results show that the GSSL 2.0 contains useful information for TL. The estimation of SOC with ≈100 transfer instances selected with ≤ 30 local data improved compared to local modelling with an equivalent number of local data and the transfer produced more stable estimates. We also showed that TL with RS-LOCAL 2.0 can be explained. The method helped to 'learn' from the specific soil information contained in the GSSL 2.0 and helped to improve the accuracy of the local estimation of SOC. The transfer relied on the spectra-SOC relationship to align the marginal and conditional distributions in the transferred data from the GSSL 2.0 and the local data. The application and further development of transfer learning in soil spectroscopy will benefit from the further development and expansion of global SSLs like the GSSL 2.0 to include additional data from under-represented regions worldwide. There are substantial opportunities for research and development of transfer learning for localizing soil spectroscopic modelling.

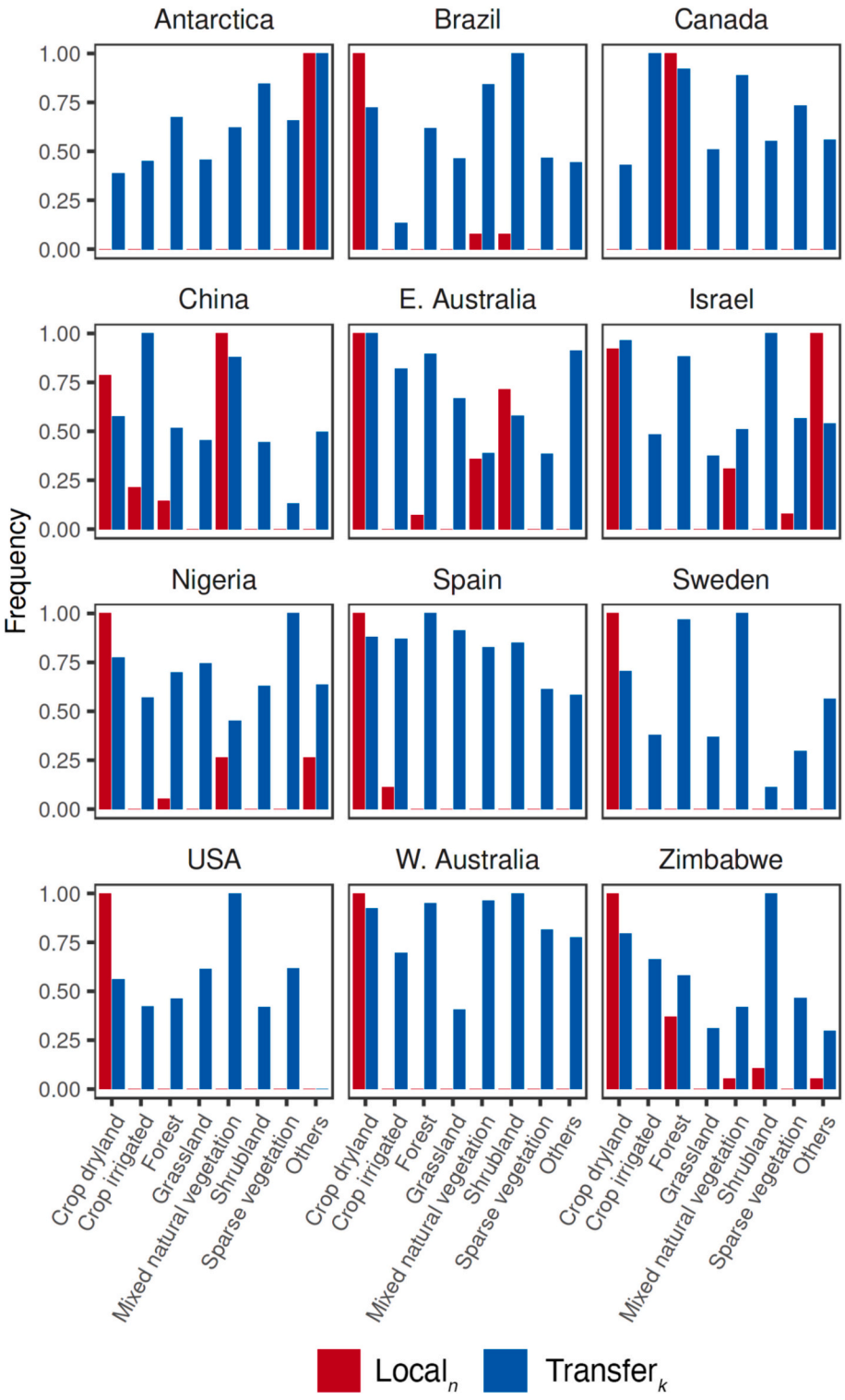

Fig. 11. Land cover type(s) of in the local$_n$ ($n = 30$) and transfer$_k$ ($k \approx 100$) data. Land cover frequencies for the local$_n$ were derived from sample counts. Because the sample counts of the land cover in the GSSL 2.0 are imbalanced, counts of land cover types in the transfer$_k$ were normalised by the total of each class in the GSSL 2.0. The largest counts in the local$_n$ and transfer$_k$ data were scaled to 1.0.

We hope that our manuscript helps to inform and incite further developments.

Data and code availability

The spectral libraries from the World Soil Information (ISRIC), the European Land Use and Coverage Area frame Survey (LUCAS) and the Mediterranean region are open source. The remaining data including the local datasets are used under agreement and are not open. They may be available from the data owners via the corresponding author on reasonable request. The RS-LOCAL 2.0 algorithm is available from the corresponding author. A version of RS-LOCAL will be available in the near future in an R software library.

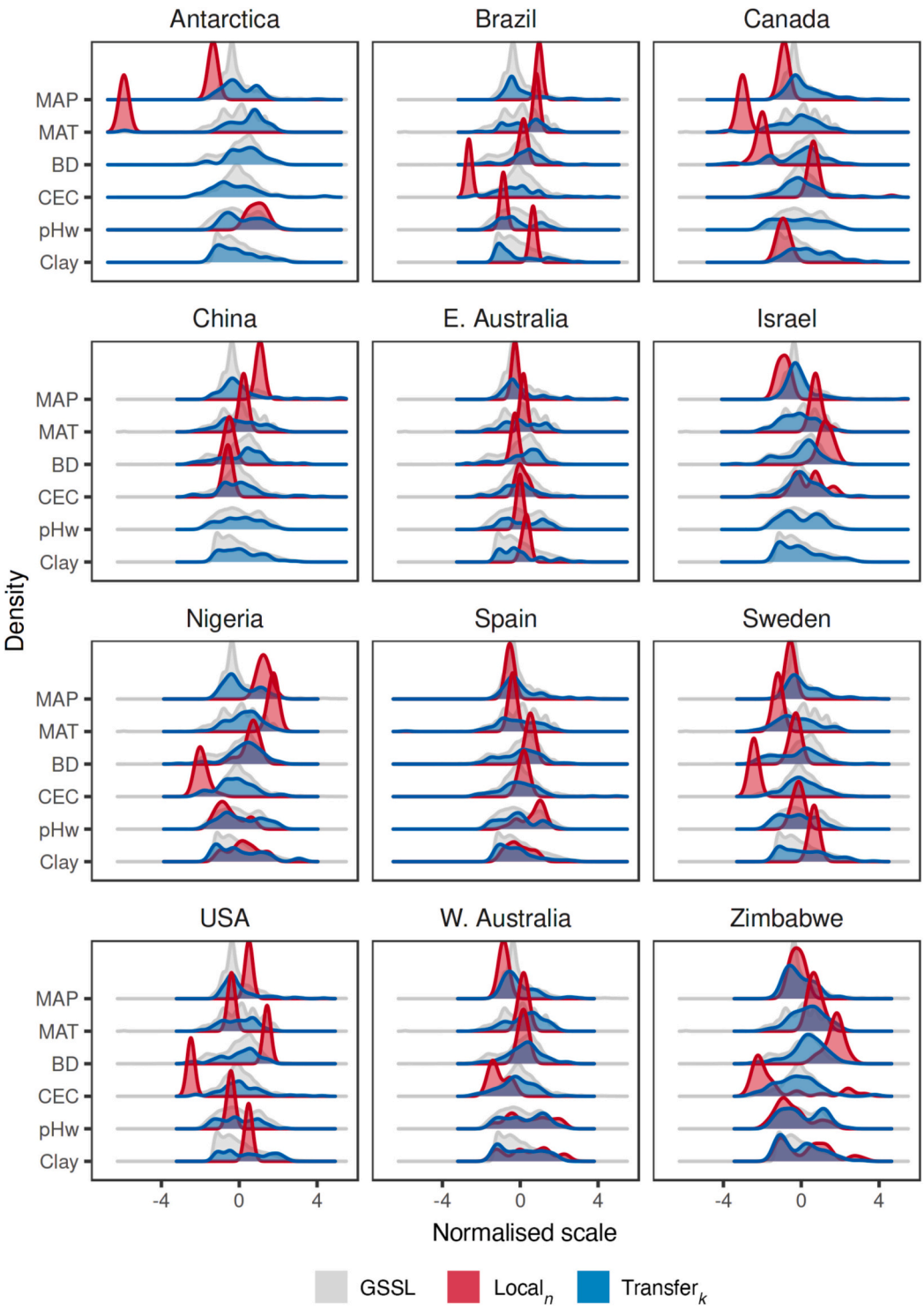

Fig. 12. Density plots showing the distribution of climate and soil properties of the GSSL 2.0, local$_n$ and transfer$_k$ data. Climate is represented by the mean annual precipitation (MAP) and mean annual temperature (MAT) and soil properties by bulk density (BD), cation exchange capacity (CEC), pH measured in water (pH_w) and clay content (Clay). Values were extracted from coarse resolution global maps.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

R.A.Viscarra Rossel reports financial support was provided by Australian Government Department of Industry Science Energy and Resources. R.A.Viscarra Rossel reports financial support was provided by Australian Research Council.

Data availability

Some data is open source, some is private, but may be available under agreement and on reasonable request.

Acknowledgements

RAVR and ZShen thank the Australian Government's Australia-China

Science and Research Fund-Joint Research Centres (ACSRF-JRCs) (grant ACSRIV000077) and RAVR thanks the Australian Research Council's Discovery Projects scheme (project DP210100420) for funding. This work was supported by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. We thank the many people who contributed to the voluntary global soil spectral library project (listed in Viscarra Rossel et al. (2016)), as well as those who contributed to the ISRIC, LUCAS, Mediterranean, and Chinese spectral libraries. We also thank Dr Mingxi Zhang who helped to source some of the global spatial datasets.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A system for Large-Scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). USENIX Association, Savannah, GA, pp. 265–283. URL. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

Andrew, A., Fearn, T., 2004. Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. Chemom. Intell. Lab. Syst. 72, 51–56. https://doi.org/10.1016/j.chemolab.2004.02.004.

Barnes, R., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43, 772–777.

Barthès, B.G., Kouakoua, E., Coll, P., Clairotte, M., Moulin, P., Saby, N.P., 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.

Baumann, P., Helfenstein, A., Gubler, A., Keller, A., Meuli, R.G., Wächter, D., Lee, J., Viscarra Rossel, R.A., Six, J., 2021. Developing the Swiss mid-infrared soil spectral library for local estimation and monitoring. Soil 7, 525–546. https://doi.org/10.5194/soil-7-525-2021.

Baumann, P., Lee, J., Behrens, T., Biswas, A., Six, J., McLachlan, G., Viscarra Rossel, R.A., 2022. Modelling soil water retention and water-holding capacity with visible–near-infrared spectra and machine learning. Eur. J. Soil Sci. 73, e13220 https://doi.org/10.1111/ejss.13220.

Behrens, T., Viscarra Rossel, R.A., Ramirez-Lopez, L., Baumann, P., 2022. Soil spectroscopy with the Gaussian pyramid scale space. Geoderma 426, 116095. https://doi.org/10.1016/J.GEODERMA.2022.116095.

Ben-Dor, E., Banin, A., 1995. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. Soil Sci. Soc. Am. J. 59, 364–372. https://doi.org/10.2136/sssaj1995.03615995005900020014x.

Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems, 24. URL. https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

Bouma, J., 2019. Soil security in sustainable development. Soil Syst. 3, 5. https://doi.org/10.3390/soilsystems3010005.

Bozinovski, S., 2020. Reminder of the first paper on transfer learning in neural networks, 1976. Informatica 44, 291–302. https://doi.org/10.31449/inf.v44i3.2828.

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

Buchhorn, M., Smets, B., Bertels, L., Roo, B.D., Lesiv, M., Tsendbazar, N.E., Herold, M., Fritz, S., 2020. Copernicus Global Land Service: Land Cover 100m: Collection 3: Epoch 2019: Globe. https://doi.org/10.5281/zenodo.3939050.

Chang, C.W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65, 480–490. https://doi.org/10.2136/sssaj2001.652480x.

Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G.A., Vergo, N., 1990. High spectral resolution reflectance spectroscopy of minerals. J. Geophys. Res. Solid Earth 95, 12653–12680. https://doi.org/10.1029/JB095iB08p12653.

Cleveland, W.S., 1981. Lowess a program for smoothing scatterplots by robust locally weighted regression. Am. Stat. 35, 54. https://doi.org/10.2307/2683591.

Cook, S.E., Jarvis, A., González, J.P., 2008. A new global demand for digital soil information. In: Digital Soil Mapping with Limited Data. Springer, Berlin, pp. 31–41.

Dai, W., Yang, Q., Xue, G.R., Yu, Y., 2007. Boosting for transfer learning. In: ACM International Conference Proceeding Series, 227, pp. 193–200. https://doi.org/10.1145/1273496.1273521.

Day, O., Khoshgoftaar, T.M., 2017. A survey on heterogeneous transfer learning. J. Big Data 4, 29. https://doi.org/10.1186/s40537-017-0089-0.

Demattê, J.A., Dotto, A.C., Paiva, A.F., Sato, M.V., Dalmolin, R.S., Maria do Socorro, B., da Silva, E.B., Nanni, M.R., ten Caten, A., Noronha, N.C., et al., 2019. The Brazilian soil spectral library (BSSL): a general view, application and challenges. Geoderma 354, 113793. https://doi.org/10.1016/j.geoderma.2019.05.043.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086.

Food and Agriculture Organization of the United Nations and Unesco, 2003. Digital soil map of the world. Version 3.6. ed. FAO, Rome. URL. https://nla.gov.au/nla.cat-vn1019159.

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

Griffiths, P., 2010. Introduction to the Theory and Instrumentation for Vibrational Spectroscopy. John Wiley & Sons, Ltd, Chichester, UK.

Guerrero, C., Zornoza, R., Gómez, I., Mataix-Beneyto, J., 2010. Spiking of NIR regional models using samples from target sites: effect of model size on prediction accuracy. Geoderma 158, 66–77. https://doi.org/10.1111/ejss.12129.

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

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A.M., Gabarrón-Galeote, M.A., Ruiz-Sinoga, J.D., Zornoza, R., Viscarra Rossel, R.A., 2016. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? Soil Tillage Res. 155, 501–509. https://doi.org/10.1016/j.still.2015.07.008.

Gupta, A., Vasava, H.B., Das, B.S., Choubey, A.K., 2018. Local modeling approaches for estimating soil properties in selected Indian soils using diffuse reflectance data over visible to near-infrared region. Geoderma 325, 59–71. https://doi.org/10.1016/j.geoderma.2018.03.025.

Helfenstein, A., Baumann, P., Viscarra Rossel, R.A., Gubler, A., Oechslin, S., Six, J., 2021. Quantifying soil carbon in temperate peatlands using a mid-IR soil spectral library. Soil 7, 193–215. https://doi.org/10.5194/soil-7-193-2021.

i-BEC, TAU, USCM, UZAY, FASF, IPB, SRTI, CUT, CEDARE, 2019. The Regional Soil Spectral Library. URL. http://datahub.geocradle.eu/dataset/regional-soil-spectral-library.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137–148.

Köchy, M., Hiederer, R., Freibauer, A., 2015. Global distribution of soil organic carbon–part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world. Soil 1, 351–365. https://doi.org/10.5194/soil-1-351-2015.

Kuhn, M., 2008. Building predictive models in R using the CARET package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05.

Lehmann, J., Bossio, D.A., Kögel-Knabner, I., Rillig, M.C., 2020. The concept and future prospects of soil health. Nat. Rev. Earth Environ. 1, 544–553. https://doi.org/10.1038/s43017-020-0080-8.

Li, S., Viscarra Rossel, R.A., Webster, R., 2022. The cost-effectiveness of reflectance spectroscopy for estimating soil organic carbon. Eur. J. Soil Sci. 73, e13202 https://doi.org/10.1111/ejss.13202.

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

Liu, L., Ji, M., Buchroithner, M., 2018. Transfer learning for soil spectroscopy based on convolutional neural networks and its application in soil clay content mapping using hyperspectral imagery. Sensors 18, 3169. https://doi.org/10.3390/s18093169.

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

Ma, C., Mu, X., Zhao, P., Yan, X., 2021. Meta-learning based on parameter transfer for few-shot classification of remote sensing scenes. Remote Sens. Lett. 12, 531–541.

Martens, H., Næs, T., 1989. Multivariate Calibration. John Wiley and Sons, Chichester, New York.

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

Mullen, K., Ardia, D., Gil, D.L., Windover, D., Cline, J., 2011. DEoptim: an R package for global optimization by differential evolution. J. Stat. Softw. 40, 1–26. URL. https://ssrn.com/abstract=1526466.

Naes, T., Isaksson, T., Kowalski, B., 1990. Locally weighted regression and scatter correction for near-infrared reflectance data. Anal. Chem. 62, 664–673.

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

Nguyen, T.T., Janik, L.J., Raupach, M., 1991. Diffuse reflectance infrared Fourier transform (DRIFT) spectroscopy in soil studies. Aust. J. Soil Res. 29, 49–67. https://doi.org/10.1071/SR9910049.

Ning, X., Wang, X., Xu, S., Cai, W., Zhang, L., Yu, L., Li, W., 2021. A review of research on co-training. In: Concurrency and Computation: Practice and Experience, 35, e6276. https://doi.org/10.1002/cpe.6276.

Niu, S., Liu, Y., Wang, J., Song, H., 2020. A decade survey of transfer learning (2010−2020). IEEE Trans. Artif. Intell. 1, 151–166. https://doi.org/10.1109/TAI.2021.3054609.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E.B., Brown, D.J., Clairotte, M., Csorba, A., et al., 2015. Soil spectroscopy: an alternative to wet chemistry for soil monitoring. Adv. Agron. 132, 139–159. https://doi.org/10.1016/bs.agron.2015.02.002.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69, 140–153. https://doi.org/10.1111/ejss.12499.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Transfer learning to localise a continental soil Vis-NIR calibration model. Geoderma 340, 279–288. https://doi.org/10.1016/j.geoderma.2019.01.009.
Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359. https://doi.org/10.1109/TKDE.2009.191.
Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q., 2011. Domain adaptation via transfer component analysis. IEEE Trans. Neural Netw. 22, 199–210. https://doi.org/10.1109/TNN.2010.2091281.
Picollo, M., Aceto, M., Vitorino, T., 2019. UV-Vis spectroscopy. Phys. Sci. Rev. 4, 20180008. https://doi.org/10.1515/psr-2018-0008.
Pittaki-Chrysodonta, Z., Hartemink, A.E., Sanderman, J., Ge, Y., Huang, J., 2021. Evaluating three calibration transfer methods for predictions of soil properties using midinfrared spectroscopy. Soil Sci. Soc. Am. J. 85, 501–519. https://doi.org/10.1002/saj2.20225.
Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7, 217–240. https://doi.org/10.5194/soil-7-217-2021.
Pratt, L.Y., Mostow, J., Kamm, C.A., 1991. Direct transfer of learned information among neural networks. In: Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2. AAAI Press, pp. 584–589.
Quinlan, J.R., 1992. Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore, pp. 343–348. https://doi.org/10.1142/9789814536271.
R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.
Rabenarivo, M., Chapuis-Lardy, L., Brunet, D., Chotte, J.L., Rabeharisoa, L., Barthès, B. G., 2013. Comparing near and mid-infrared reflectance spectroscopy for determining properties of Malagasy soils, using global or local calibration. J. Near Infrared Spectrosc. 21, 495–509. URL. https://opg.optica.org/jnirs/abstract.cfm?URI=jnirs-21-6-495.
Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T., 2013. The spectrum-based learner: a new local approach for modeling soil Vis–NIR spectra of complex datasets. Geoderma 195, 268–279. https://doi.org/10.1016/j.geoderma.2012.12.014.
Sankey, J.B., Brown, D.J., Bernard, M.L., Lawrence, R.L., 2008. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. Geoderma 148, 149–158. https://doi.org/10.1016/j.geoderma.2008.09.019.
Seidel, M., Hutengs, C., Ludwig, B., Thiele-Bruhn, S., Vohland, M., 2019. Strategies for the efficient estimation of soil organic carbon at the field scale with Vis-NIR spectroscopy: spectral libraries and spiking vs. local calibrations. Geoderma 354, 113856. https://doi.org/10.1016/j.geoderma.2019.07.014.
Shen, Z., Viscarra Rossel, R.A., 2021. Automated spectroscopic modelling with optimised convolutional neural networks. Sci. Rep. 11, 1–12. https://doi.org/10.1038/s41598-020-80486-9.
Shen, Z., Ramirez-Lopez, L., Behrens, T., Cui, L., Zhang, M., Walden, L., Wetterlind, J., Shi, Z., Sudduth, K.A., Baumann, P., Song, Y., Catambay, K., Viscarra Rossel, R.A., 2022. Deep transfer learning of global spectra for local soil carbon monitoring. ISPRS J. Photogramm. Remote Sens. 188, 190–200. https://doi.org/10.1016/j.isprsjprs.2022.04.009.
Shenk, J.S., Westerhaus, M.O., Berzaghi, P., 1997. Investigation of a LOCAL calibration procedure for near infrared instruments. J. Near Infrared Spectrosc. 5, 223–232.
Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66, 988–998. https://doi.org/10.2136/sssaj2002.9880.
Shepherd, K.D., Palm, C.A., Gachengo, C.N., Vanlauwe, B., 2003. Rapid characterization of organic resource quality for soil and livestock management in tropical agroecosystems using near-infrared spectroscopy. Agron. J. 95, 1314–1322. https://doi.org/10.2134/agronj2003.1314.
Shepherd, K.D., Ferguson, R., Hoover, D., van Egmond, F., Sanderman, J., Ge, Y., 2022. A global soil spectral calibration library and estimation service. Soil Secur. 7, 100061. https://doi.org/10.1016/j.soisec.2022.100061.
Shi, Z., Ji, W., Viscarra Rossel, R.A., Chen, S., Zhou, Y., 2015. Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese Vis–NIR spectral library. Eur. J. Soil Sci. 66, 679–687. https://doi.org/10.1111/ejss.12272.
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, 139–186. https://doi.org/10.1080/05704928.2013.811081.
St. Luce, M.S., Ziadi, N., Viscarra Rossel, R.A., 2022. GLOBAL-LOCAL: a new approach for local predictions of soil organic carbon content using large soil spectral libraries. Geoderma 425, 116048. https://doi.org/10.1016/j.geoderma.2022.116048.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. Adv. Agron. 107, 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7.
Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PLoS One 8, 1–13. https://doi.org/10.1371/journal.pone.0066409.
Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A survey on deep transfer learning. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27. Springer, pp. 270–279.
Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B., Zalidis, G.C., 2020. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. Geoderma 367, 114208. https://doi.org/10.1016/j.geoderma.2020.114208.
Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T., 2014. Deep domain confusion: Maximizing for domain invariance arXiv preprint arXiv:1412.3474.
Tziolas, N., Tsakiridis, N., Ben-Dor, E., Theocharis, J., Zalidis, G., 2019. A memory-based learning approach utilizing combined spectral sources and geographical proximity for improved VIS-NIR-SWIR soil properties estimation. Geoderma 340, 11–24. https://doi.org/10.1016/j.geoderma.2018.12.044.
Vapnik, V., Golowich, S., Smola, A., 1996. Support vector method for function approximation, regression estimation and signal processing. Adv. Neural Inf. Proces. Syst. 9, 281–287. URL. https://proceedings.neurips.cc/paper_files/paper/1996/file/4f284803bd0966cc24fa8683a34afc6e-Paper.pdf.
Vasques, G.M., Grunwald, S., Harris, W.G., 2010. Spectroscopic models of soil organic carbon in Florida, USA. J. Environ. Qual. 39, 923–934. https://doi.org/10.2134/jeq2009.0314.
Viscarra Rossel, R.A., 2007. Robust modelling of soil diffuse reflectance spectra by 'bagging-partial least squares regression'. J. Near Infrared Spectrosc. 15, 39–47. https://doi.org/10.1255/jnirs.694.
Viscarra Rossel, R.A., 2011. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. J. Geophys. Res. 116, F04023 https://doi.org/10.1029/2011JF001977.
Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025.
Viscarra Rossel, R.A., Bouma, J., 2016. Soil sensing: a new paradigm for agriculture. Agric. Syst. 148, 71–74. https://doi.org/10.1016/j.agsy.2016.07.001.
Viscarra Rossel, R.A., Hicks, W.S., 2015. Soil organic carbon and its fractions estimated by visible–near infrared transfer functions. Eur. J. Soil Sci. 66, 438–450. https://doi.org/10.1111/ejss.12237.
Viscarra Rossel, R.A., Lark, R.M., 2009. Improved analysis and modelling of soil diffuse reflectance spectra using wavelets. Eur. J. Soil Sci. 60. https://doi.org/10.1111/j.1365-2389.2009.01121.x.
Viscarra Rossel, R.A., McBratney, A., 1998. Soil chemical analytical accuracy and costs: implications from precision agriculture. Aust. J. Exp. Agric. 38, 765–775. https://doi.org/10.1071/EA97158.
Viscarra Rossel, R.A., Webster, R., 2012. Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. Eur. J. Soil Sci. 63, 848–860. https://doi.org/10.1111/j.1365-2389.2012.01495.x.
Viscarra Rossel, R.A., Cattle, S.R., Ortega, A., Fouad, Y., 2009. In situ measurements of soil colour, mineral composition and clay content by Vis–NIR spectroscopy. Geoderma 150, 253–266. https://doi.org/10.1016/j.geoderma.2009.01.025.
Viscarra Rossel, R.A., Bui, E., De Caritat, P., McKenzie, N., 2010. Mapping iron oxides and the color of Australian soil using visible-near-infrared reflectance spectra. J. Geophys. Res. Earth 115, F04031. https://doi.org/10.1029/2009JF001645.
Viscarra Rossel, R.A., Webster, R., Bui, E., Baldock, J., 2014. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Glob. Chang. Biol. 20, 2953–2970. https://doi.org/10.1111/gcb.12569.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J., Shepherd, K., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aichi, H., Barthes, B., Bartholomeus, H., Bayer, A., Bernoux, M., Bottcher, K., Brodsky, L., Du, C., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C., Knadel, M., Morras, H., Nocita, M., Ramírez López, L., Roudier, P., Campos, E., Sanborn, P., Sellitto, V., Sudduth, K., Rawlins, B., Walter, C., Winowiecki, L., Hong, S., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth Sci. Rev. 155, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012.
Viscarra Rossel, R.A., Lee, J., Behrens, T., Luo, Z., Baldock, J., Richards, A., 2019. Continental-scale soil carbon composition and vulnerability modulated by regional environmental controls. Nat. Geosci. 12, 547–552. https://doi.org/10.1038/s41561-019-0373-z.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Ge, Y., Gomez, C., Guerrero, C., Peng, Y., Ramirez-Lopez, L., Shi, Z., Stenberg, B., Webster, R., Winowiecki, L., Shen, Z., 2022. Diffuse reflectance spectroscopy for estimating soil properties: a technology for the 21st century. Eur. J. Soil Sci. 73, e13271 https://doi.org/10.1111/ejss.13271.
Vohland, M., Ludwig, M., Harbich, M., Emmerling, C., Thiele-Bruhn, S., 2016. Using variable selection and wavelets to exploit the full potential of visible–near infrared spectra for predicting soil properties. J. Near Infrared Spectrosc. 24, 255–269.
Wang, Z., Dai, Z., Póczos, B., Carbonell, J., 2019. Characterizing and avoiding negative transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, California, pp. 11293–11302.
Weiss, K., Khoshgoftaar, T.M., Wang, D.D., 2016. A survey of transfer learning. J. Big Data 3, 1–40. https://doi.org/10.1186/S40537-016-0043-6/TABLES/6.
Wetterlind, J., Stenberg, B., 2010. Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. Eur. J. Soil Sci. 61, 823–843. https://doi.org/10.1111/j.1365-2389.2010.01283.x.
Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S., West, L.T., 2014. Overview of the US rapid carbon assessment project: sampling design, initial summary and uncertainty estimates. Soil Carbon 95–104.
Wold, S., Sjostrom, M., Eriksson, L., 2001. PLS-Regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

World Agroforestry (ICRAF), International Soil Reference Information Centre (ISRIC), 2021. ICRAF-ISRIC Soil VNIR Spectral Library. https://doi.org/10.34725/DVN/MFHA9C.

Xu, S., Shi, X., Wang, M., Zhao, Y., 2016. Effects of subsetting by parent materials on prediction of soil organic matter content in a hilly area using Vis–NIR spectroscopy. PLoS One 11, e0151536. https://doi.org/10.1371/journal.pone.0151536.

Yang, Y., Viscarra Rossel, R.A., Li, S., Bissett, A., Lee, J., Shi, Z., Behrens, T., Court, L., 2019. Soil bacterial abundance and diversity better explained and predicted with spectro-transfer functions. Soil Biol. Biochem. 129, 29–38. https://doi.org/10.1016/j.soilbio.2018.11.005.

Yang, Y., Shen, Z., Bissett, A., Viscarra Rossel, R.A., 2022. Estimating soil fungal abundance and diversity at a macroecological scale with deep learning spectrotransfer functions. Soil 8, 223–235. https://doi.org/10.5194/soil-8-223-2022.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks?. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. URL. https://proceedings.neurips.cc/paper_files/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Springer, Zurich, Switzerland, pp. 818–833.

Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L., 2019. S4L: Self-supervised semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2020. A comprehensive survey on transfer learning. Proc. IEEE 109, 43–76.