https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

SOIL Discussions

Estimating soil fungal abundance and diversity at a macroecological scale with deep learning spectrotransfer functions

Yuanyuan Yang[1], Zefang Shen[1], Andrew Bisset[2], and Raphael A. Viscarra Rossel[1]

[1]Soil and Landscape Science, School of Molecular and Life Sciences, Curtin University, GPO Box U1987, Perth WA 6845, Australia.
[2]CSIRO Oceans and Atmosphere, GPO BOX 1538, Hobart TAS 7001, Australia.

Correspondence: Raphael A. Viscarra Rossel (r.viscarra-rossel@curtin.edu.au)

Abstract. Soil fungi play important roles in the functioning of ecosystems, but they are challenging to measure. Using a continental scale dataset, we developed and evaluated a new method to estimate the relative abundance of the dominant phyla and diversity of fungi in Australian soil. The method relies on the development of spectro-transfer functions with state-of-the-art machine learning and using publicly available data on soil and environmental proxies for edaphic, climatic, biotic and topographic factors, and visible–near infrared (vis–NIR) wavelengths, to estimate the relative abundances of the Ascomycota, Basidiomycota, Glomeromycota, Mortierellomycota and Mucoromycota and community diversity measured with the abundance-based coverage estimator (ACE) index. The machine learning algorithms tested were partial least squares regression (PLSR), random forest (RF), Cubist, support vector machines (SVM), Gaussian process regression (GPR), XG-boost (XGB) and one-dimensional convolutional neural networks (1D-CNNs). The spectro-transfer functions were validated with a 10-fold cross-validation (n = 577). The 1D-CNNs outperformed the other algorithms and could explain between 45 and 73 % of fungal relative abundance and diversity. The models were interpretable, and showed that soil nutrients, pH, bulk density, an ecosystem water balance (a proxy for aridity) and net primary productivity were important predictors, as were specific vis–NIR wavelengths that correspond to organic functional groups, iron oxide and clay minerals. Estimates of the relative abundance for Mortierellomycota and Mucoromycota produced $R^2 \geq 0.60$, while estimates of the abundance of the Ascomycota and Basidiomycota produced R^2 values of 0.5 and 0.58, respectively. The spectro-transfer functions for the Glomeromycota and diversity were the poorest with R^2 values of 0.48 and 0.45, respectively. There is no doubt that the method provides estimates that are less accurate than more direct measurements with conventional molecular approaches. However, once the spectro-transfer functions are developed, they can be used with very little cost, and could serve to supplement the more expensive and laborious molecular approaches for a better understanding of soil fungal abundance and diversity under different agronomic and ecological settings.

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

1 Introduction

Soil fungi are important components of microbial communities, which inhabit dynamic soil environments. They play critical functional roles as decomposers, mutualists, and pathogens (Li et al., 2019). They impact nutrient cycling and ecosystem services, such as soil carbon fixation, fertility and productivity (Vetrovsky et al., 2019; Delgadobaquerizo et al., 2016). Given the important functions that soil fungi perform, it is important to better characterise and understand their communities over large scales. However, data on soil fungi are few or largely unavailable because the measurement of soil fungi, which needs field sampling, followed by culture-based analysis or DNA sequencing, are laborious, time-consuming and costly. Using soil sensing technologies, such as spectroscopy together with molecular approaches could greatly improve the utility of fungal inventory data (Hart et al., 2020).

Improvements in soil analytical methodologies provide an opportunity to increase sampling density for deriving a more detailed understanding of soil properties, their spatial variation, soil condition, and to improve decision-making. Spectroscopic techniques, such as visible–near infrared (vis–NIR) spectroscopy, have been developed to provide rapid estimates of soil properties (Viscarra Rossel et al., 2016). Soil vis–NIR spectra are largely nonspecific because of the overlapping absorptions of soil constituents (Stenberg et al., 2010). Complex absorption patterns generated from soil constituents need to be mathematically extracted from the spectra and there are various methods that can be used to model soil properties with spectra. They include multivariate statistical methods such as partial least squares regression (PLSR), and machine learning with different algorithms, including neural networks (Viscarra Rossel and Behrens, 2010; Morellos et al., 2016; Liu et al., 2018; Tsakiridis et al., 2020; Shen and Viscarra Rossel, 2021). Thus, vis–NIR spectra can integrally characterize the soil's mineral-organic composition, and combined with multivariate modelling, soil spectroscopy provides a rapid and cost-efficient method for soil characterisation (Viscarra Rossel and Brus, 2018).

Although there are no vis–NIR absorptions that can be directly assigned to soil microbial communities or diversity, soil microbes are dependent on the fundamental soil composition: its minerals, organic matter and water content. For example, they rely on organic matter for energy, on clay minerals and iron oxides for the supply of essential elements to grow (Müller, 2015). These organic and mineral properties are well represented and have a direct response in soil vis–NIR spectra (Stenberg et al., 2010). Therefore, vis–NIR spectra have been used to model various functional soil properties, such as soil organic carbon, cation exchange capacity, pH, clay content (Shi et al., 2015), as well as soil microbial communities (Davinic et al., 2012; Yang et al., 2019). For the latter, if the microbial biomass is present in the soil organic matter, then the spectra might well detect their functional constituents.

There are studies that use environmental proxies, (or covariates) at continental and global scales to model soil microbial properties using various methods, including linear regressions and machine learning (Serna-Chavez et al., 2013; Griffiths et al., 2011; Vetrovsky et al., 2019; Yang et al., 2019; Delgadobaquerizo et al., 2018a). However, we found no published studies that used vis–NIR spectra or a combination of spectra with other soil and environmental covariates (i.e. spectro-transfer functions) to infer fungal abundance or diversity. In a previous study, Yang et al. (2019) showed that vis–NIR spectra combined with other soil and environmental data could estimate soil bacterial abundance and diversity. Here, our hypotheses are: (i) spectroscopic

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

models with machine learning can estimate soil fungal abundance and diversity at the continental scale, and (ii) spectro-transfer functions with additional predictors to capture other soil and environmental properties that affect soil fungi will improve the accuracy of the estimates.

Thus, our objective is to develop and test the spectroscopic method for estimating soil fungal abundance and diversity over a large scale, and our aims are to:

(i) Compare the modelling of fungal abundance and diversity with vis–NIR spectra only (spectroscopic models), with readily available soil and environmental data only (environmental models) and with the combined set of vis–NIR spectra and readily available soil and environmental data (spectro-transfer functions), and

(ii) Test different statistical and machine learning algorithms for the modelling.

2 Methods

2.1 Soil sampling and laboratory analyses

We used 577 soil samples from the Biomes of Australian Soil Environments (BASE) project (Bissett et al., 2016). In that project, the sampling was undertaken from soil that supports diverse plant communities across Australia. Samples came from two soil depths (0–0.1m and 0.2–0.3m), covering five typical Australian ecosystem types, including cropland, forest, grassland, shrubland, and woodland (Fig. 1a). Each sample was partitioned into subsamples for DNA sequencing (see below) and air-dried and crushed to a particle size of $\leq$2 mm for physicochemical analyses. The soil properties analyzed were total organic carbon and soil nutrients (e.g. ammonium, nitrate, phosphorus, potassium), pH, exchangeable cations (aluminium, sodium, magnesium, calcium), and texture (sand, silt and clay). The methods are described in (Bissett et al., 2016). Subsamples of the $\leq$2 mm portions were used for the spectroscopic analysis (see below).

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

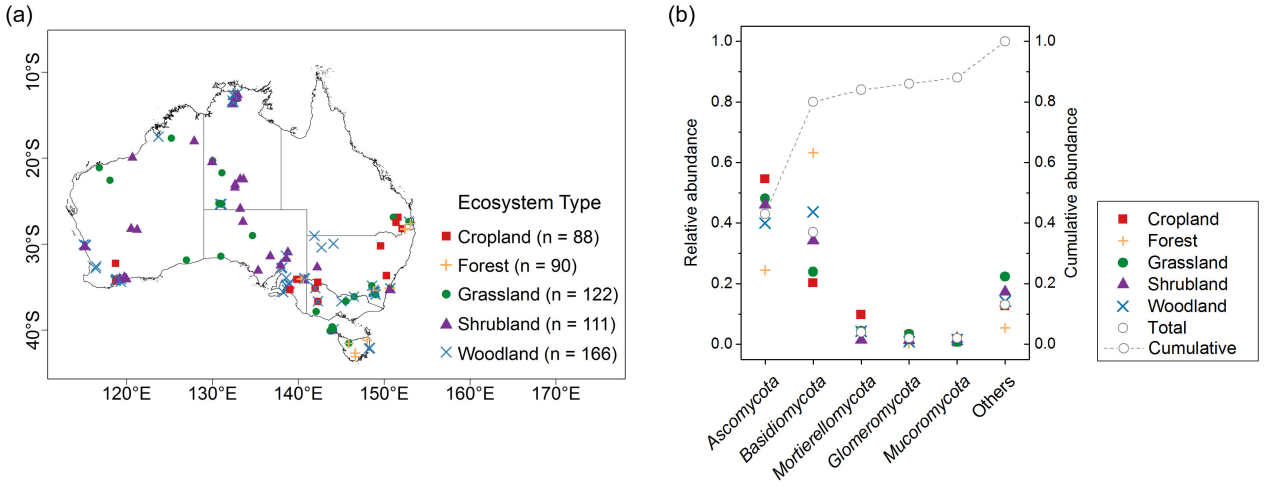

Figure 1. (a) Sampling sites and the range of ecosystem types across Australia (b) Relative abundances of dominant fungal phyla and unclassified "Others" taxa in five ecosystem types. Individual abundance of each phylum and their cumulative abundance were shown in the graph.

2.2 Derivation of fungal abundance and diversity

The methods for DNA extraction and sequencing are detailed in Bissett et al. (2016). Briefly, the soil DNA was extracted in triplicate following methods used in the Earth Microbiome Project[1]. Sequencing occurred with an Illumina MiSEQ, which is described in the BASE protocols[2]. Summarising, amplicons targeting the fungal ITS region were prepared and sequenced for each sample. The ITS amplicons were sequenced using 300 bp paired end sequencing. ITS1 regions were extracted using ITSx Bengtsson-Palme et al. (2013). Sequences comprising full and partial ITS1 regions were passed to the Operational Taxonomic Units (OTU) selection and assigning workflow Bissett et al. (2016), which followed guidelines described in the BASE protocols[3] and in Bissett et al. (2016). These are based on the most current version of UNITE database (version 8.2, updated 15-01-2020) for molecular identification of fungi Nilsson et al. (2018). We used the final sample-by-OTU data matrix and annotated taxonomy file for the analyses of fungal diversity and composition.

In total, there were more than 60 million quality sequences across the samples, with 11,090–2,177,737 sequences per sample (mean 107,310). Sequences clustered into 202,200 OTUs at 97% similarity, with an average of 666 OTUs per sample. To eliminate bias on the diversity comparison caused by unbalanced sequencing, samples were resampled at the same sequencing depth using functions of the RAM library in the R software(R Core Team, 2014). Here, 11 000 sequences (the median number of sequences in the samples) were used as the resampling depth, because the majority of samples only had this amount of sequences, but also because the rarefaction curves started to flatten out for all 577 samples at this sequencing depth. This

[1]http://www.Earthmicrobiome.Org/emp-standard-protocols/dna-extraction-protocol/
[2]https://ccgapps.Com.Au/bpa-metadata/base/information
[3]https://ccgapps.com.au/bpa-metadata/base/information

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

suggested that the sequencing number was sufficient (Fig. S1 in the Supplementary Information). To quantify community diversity, we then calculated the abundance-based coverage estimator (ACE) index (Lozupone and Knight, 2008) from the resampled sample-by-OTU matrix. The relative abundance of fungal phyla were then determined using the ratio of sequences number classified at each phylum to the total number of sequences of each sample.

2.3 Soil visible–near-infrared spectroscopy

We measured the diffuse reflectance spectra of all air-dried $\leq$ 2 mm soil samples with the Labspec® vis–NIR spectrometer (Malvern Panalytical, Boulder, Colorado, USA) following the protocols described in Viscarra Rossel et al. (2016). The spectral range of the spectrometer is 350 to 2500 nm. Due to the low signal-to-noise ratio at the start and end of each spectrum, for our analysis, we kept only spectra in the range between 380 and 2450 nm. As the spectra are highly collinear, to reduce redundancy in the data, we re-sampled them to a resolution of 10 nm. The measurements were performed with the instruments high intensity contact probe (PaNalytic, Boulder, Colorado, USA), and a Spectralon® white reference panel was used for calibration once every 10 measurements.

To interpret the spectra, we fitted each reflectance (R) spectrum with a convex hull and computed the deviations from the hull (Clark and Roush, 1984). These continuum removed (CR) spectra help to identify characteristic absorptions of soil constituents. For the modelling, we first transformed the R spectra to apparent absorbance, using $\mathrm{A} = \log_{10}(1/\mathrm{R})$, and then used a Savitzky-Golay filter with a window of size 7 and a fitting polynomial of degree 2 and first derivative Savitzky and Golay (1964) to remove baseline effects and to improve the signal to noise ratio.

2.4 Modelling soil fungal abundance and diversity

We developed spectroscopic models, environmental models, and spectro-transfer functions for estimating soil fungal abundance and diversity (see below). The spectroscopic models used only the vis–NIR spectra, the environmental models used only the publicly available soil and environmental data that represent the soil forming factors soil, climate, vegetation, terrain and parent material (Jenny, 1994), and the spectro-transfer functions used the vis—NIR spectra together with soil and environmental data.

We assembled a set of readily available soil and environmental maps that represented climate, terrain, vegetation, and parent material. To relate the these covariates to the fungal data, we extracted values from these maps using the geographic coordinates of the sample set. The soil property data came from the Australia-wide fine spatial resolution (90 $\times$ 90 m) digital soil maps of total organic carbon, total nitrogen, total phosphorus, bulk density, effective cation exchange capacity, available water capacity, pH, and soil texture (sand, silt, and clay) (Viscarra Rossel et al., 2015), as well as maps of the clay minerals kaolinite, illite, and smectite (Viscarra Rossel, 2011). To represent climate, we used data on mean annual temperature (MAT), mean annual precipitation (MAP), solar radiation, and evapotranspiration (Xu and Hutchinson, 2011) and the Prescott index (PI) (Prescott, 1950), which is calculated as the ratio of precipitation to evapotranspiration. To capture functional landscape characteristics, we used a digital elevation model (DEM) from the 3-arc second shuttle radar topographic mission (SRTM) and derived terrain attributes (Gallant et al., 2011). To represent vegetation, we used data on net primary productivity (NPP) (Haverd et al., 2013), and on the fraction of photosynthetically active radiation intercepted by the sunlit canopy of the evergreen (Fpar-e) and woody

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

(Fpar-r) vegetation (Donohue et al., 2009). To represent parent material, we used gamma radiometrics, which comprises data on potassium, uranium, and thorium (Minty et al., 2009). Supplementary Table S1 lists these data and their main characteristics.

The spectra and the covariates were centred and scaled before the modelling of fungal abundance and diversity. The algorithms that we tested were partial least squares regression (PLSR) (Wold et al., 2001), gaussian process regression(GPR) (Rasmussen and Williams, 2005), support vector machines (SVM) (Suykens et al., 2002), random forest(RF) (Breiman, 2001a), CUBIST (Quinlan, 1992), extreme gradient boost (XGBoost) (Friedman, 2001) and optimised 1D convolutional neural networks (1D-CNNs) (Shen and Viscarra Rossel, 2021). The algorithms and their implementation are described in the Supplementary Information linked to this article.

The predictability of the spectroscopic models and the spectro-transfer functions were assessed using 10-fold cross-validations. We evaluated the estimates using the coefficient of determination (R^2), the root mean squared error (RMSE), which measures inaccuracy, the standard deviation of the error (SDE), which measures imprecision and the mean error (ME), which measures bias (Viscarra Rossel and McBratney, 1998). Inaccuracy (RMSE) embraces both the bias (ME) and the imprecision (SDE) (Viscarra Rossel and McBratney, 1998). Their relationship is given by $\mathrm{RMSE}^2 = \mathrm{ME}^2 + \mathrm{SDE}^2$.

To interpret the models, we calculated their variable importance as follows. For the PLSR, GPR, SVM, Cubist, RF and XGBoost models, variable importance was calculated using the varImp function in the caret library (Kuhn et al., 2008) of the software R. For the 1D-CNNs, we used the permutation importance (Breiman, 2001b; Fisher et al., 2019). In the results, we only report the variable importance of the model that performed best.

3 Results

In total, more than 60 million quality filtered sequences in the whole dataset were obtained, with an average of 107 310 sequences per sample. When we clustered the sequences at 97% similarity level 202 200 OTUs were detected. Each sample had an average of 666 OTUs. Sixteen phyla were identified in total and 5 dominant phyla, with relative abundance $> 2\%$, were approximately present in most soils. This represented nearly 88% of the sequence number. The relative abundance of fungal phyla varied across ecosystem types (Fig. 1b).

Ascomycota (mean 0.43, SD 0.21) was the most abundant phylum, followed by Basidiomycota (mean 0.37, SD 0.24) (Table 1). Dominant fungal phyla showed a high degree of variability, with an averaging 83% coefficient of variation (CV). The ACE index showed a wide range from 81 to 1823 (mean 563, SD 315). The rich soil biodiversity of the data resulted from the extensitive soil sampling taken from diverse vegetation, soils, and climates across Australia.

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

Table 1. The descriptive statistics of relative abundance of dominant phyla and community diversity (n = 577).

Variables	Mean	Median	St. Dev.	Range	Coeff. var. (%)
Abundance					
Ascomycota	0.43	0.42	0.21	0.04–0.98	49
Basidiomycota	0.37	0.32	0.24	0.01–0.92	65
Mortierellomycota	0.04	0.02	0.04	0.00–0.36	100
Glomeromycota	0.02	0.01	0.01	0.00–0.41	50
Mucoromycota	0.02	0.01	0.03	0.00–0.55	150
Diversity					
ACE	563	503	315	81–1823	56

Fig. 2 shows the CR spectra with the characteristic absorptions. Soil with different fungal diversity show variations in absorptions, particularly around those that are due to Fe-oxides (400–800 nm), minerals (around 1400 nm, 1900 nm and 2200 nm) and organic compounds (throughout the vis–NIR spectrum) (Stenberg et al., 2010). Soil with the lower fungal diversity showed a more pronounced absorbance around 600 nm as shown in Fig. 2. In our study, the soil with lower fungal
155 diversity mainly come from the central and western Australia. In these areas, soil subjected to intense weathering regimes and can accumulate large quantities Fe oxides (total soil Fe_2O_3 larger than 10%) in surficial environments, and strongly absorbed in the visible region (Viscarra Rossel et al., 2010). These highly iron-rich lateritic soil occur with acidic pH, high H_2O and Al activities, and has been shown not conductive to the development of fungal diversity (Viscarra Rossel et al., 2010).

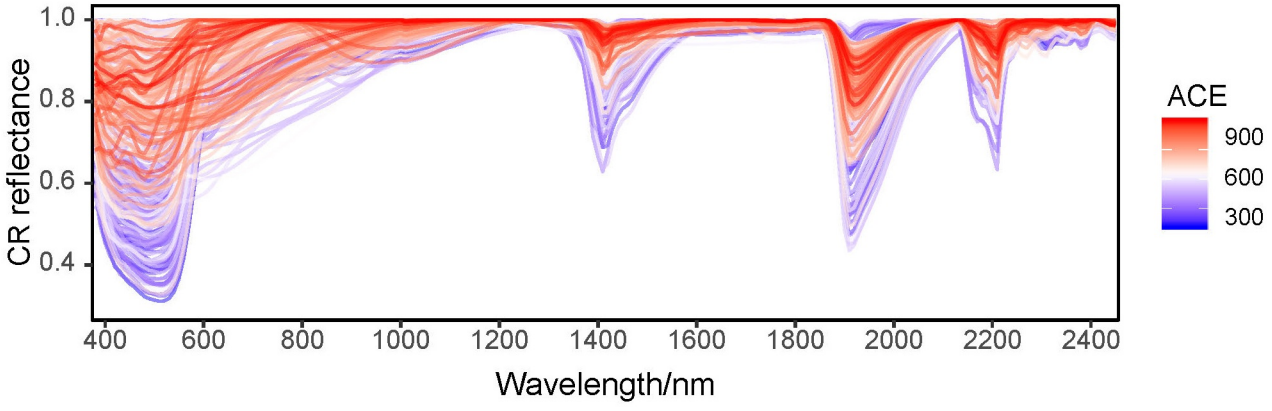

Figure 2. Continuum removed (CR) spectra curves colored by fungal ACE diversity

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

3.1 Modelling

With the different algorithms, the spectroscopic models (i.e. with only the vis–NIR spectra) could explain 9–45% of the variation in fungal phyla relative abundance and diversity. Spectroscopic models of the Glomeromycota were the least successful, with R^2 values ranging from 0.09 using SVM to 0.30 using 1D-CNN, while those of the Mortierellomycota produced the largest R^2 values, ranging from 0.32 using XGBoost to 0.45 using 1D-CNN (Fig. 3). The models of diversity had R^2 values ranging from 0.14 with PLSR to 0.35 using 1D-CNN.

The models derived with the readily available soil and environment data could explain 14–60% of the variation in fungal phyla relative abundance and diversity with the different algorithms. These environmental models were generally better performed than spectroscopic models, with an average 10% additional variance explained.

Combining the vis–NIR spectra and soil and environmental data further improved the modelling and their explanatory power. The spectro-transfer functions (i.e. with the combined set of vis–NIR spectra and other soil and environmental data) performed, on average, 20% better than the spectroscopic models and 10% better than environmental models. Depending on the algorithm used, they could explain between 17–73% of the variation in fungal phyla relative abundance and diversity (Fig. 3). The spectro-transfer functions of Glomeromycota produced R^2 values ranging from 0.17 using PLSR to 0.48 using 1D-CNN. The spectro-transfer functions of the Mortierellomycota and Mucoromycota produced the largest R^2 values ranging from 0.51 to 0.73 (Fig. 3).

Generally, PLSR and GPR were the least successful methods, while SVM, RF, Cubist and XGBoost were similarly successful for estimating fungal phyla relative abundance and diversity (Fig. 3). The 1D-CNN spectro-transfer functions were 13–31% more successful compared to other machine learning methods as they could explain between 45–73% of the variation in fungal relative abundance and diversity (Fig. 3).

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

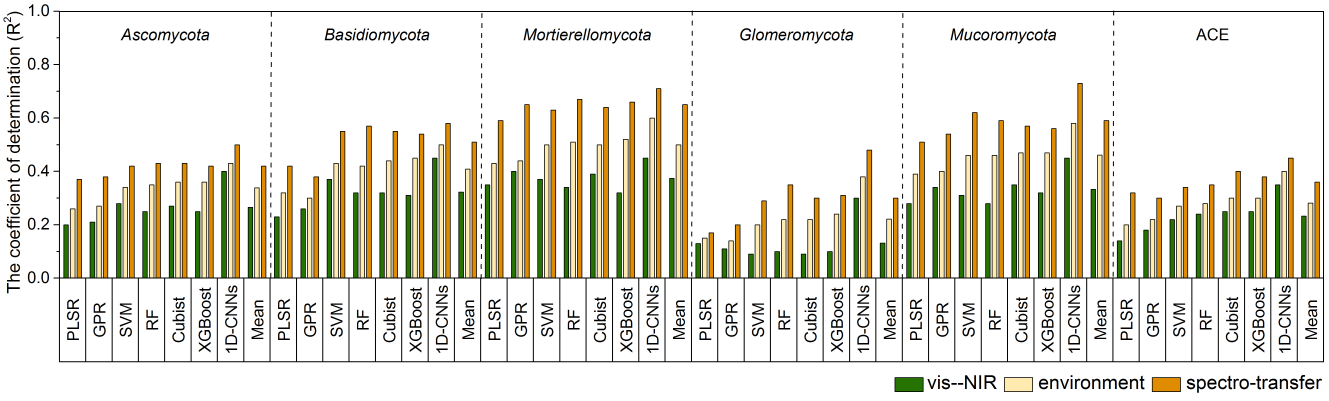

Figure 3. The coefficient of determination (R^2) for the vis–NIR spectroscopic models, soil and environmental models and the spectro-trasnfer functions that used combined set of the vis–NIR and readily available soil and environmental covariates, to estimate soil fungal phyla abundance and diversity (n = 577). The different statistical and machine learning methods were partial least squares regression (PLSR), gaussian process regression (GPR), support vector machines (SVM), random forest(RF), CUBIST, extreme gradient boost (XGBoost) and optimised 1D convolutional neural networks (1D-CNNs).

3.2 1D-CNNs spectro-transfer functions

The final architectures and optimised hyperparameters of the 1D-CNNs are given in Supplementary Table S3. As deep learning models are dataset dependent, the optimisation returned a different architecture for each response variable. Overall, the 1D-CNNs used simple architectures with less than 4 convolutional layers (Supplementary Table S3). Scatter plots of the measured versus estimated values of relative abundance and diversity using 1D-CNNs spectro-transfer functions and their validation statistics are shown in Fig. 4. Estimates of the relative abundance of Mortierellomycota and Mucoromycota produced R^2 values ≥ 0.60, while estimates of Ascomycota and Basidiomycota produced $0.5 \leq R^2 < 0.6$. Estimates of Glomeromycota and ACE produced $0.4 \leq R^2 < 0.5$. The estimates were relatively unbiased (small ME), although generally small values were overestimated and large values were underestimated (Fig. 4). Imprecision contributed to the majority of the RMSE. The imprecision of our estimates was a result of absence of repeated sampling and high adaptability of soil fungi to the wide range of environments.

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

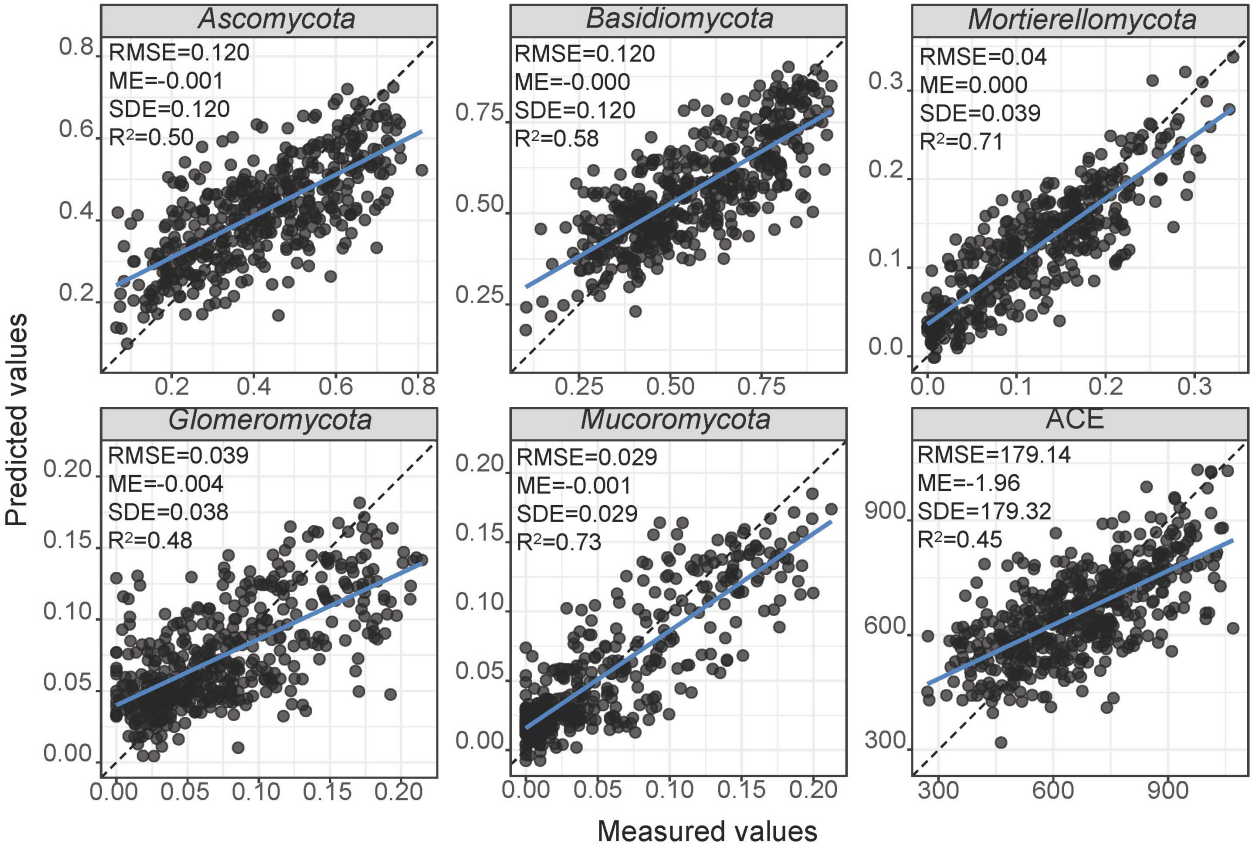

Figure 4. Performance of the CNN spectro-transfer functions for estimate of the relative abundance of dominant fungal phyla and diversity index. The spectro-transfer functions used vis–NIR spectra with other publicly available data on soil environmental variables. The plots show measured vs. estimated values using a 10-folds cross validation. The gray points represent no overlap with any other points, and the black points represent at least two points that overlap.

190 The important variables in the 1D-CNNs spectro-transfer functions of phyla relative abundance and diversity were vis–NIR wavelengths representing organic matter, iron oxide and clay minerals (Fig. 5).

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

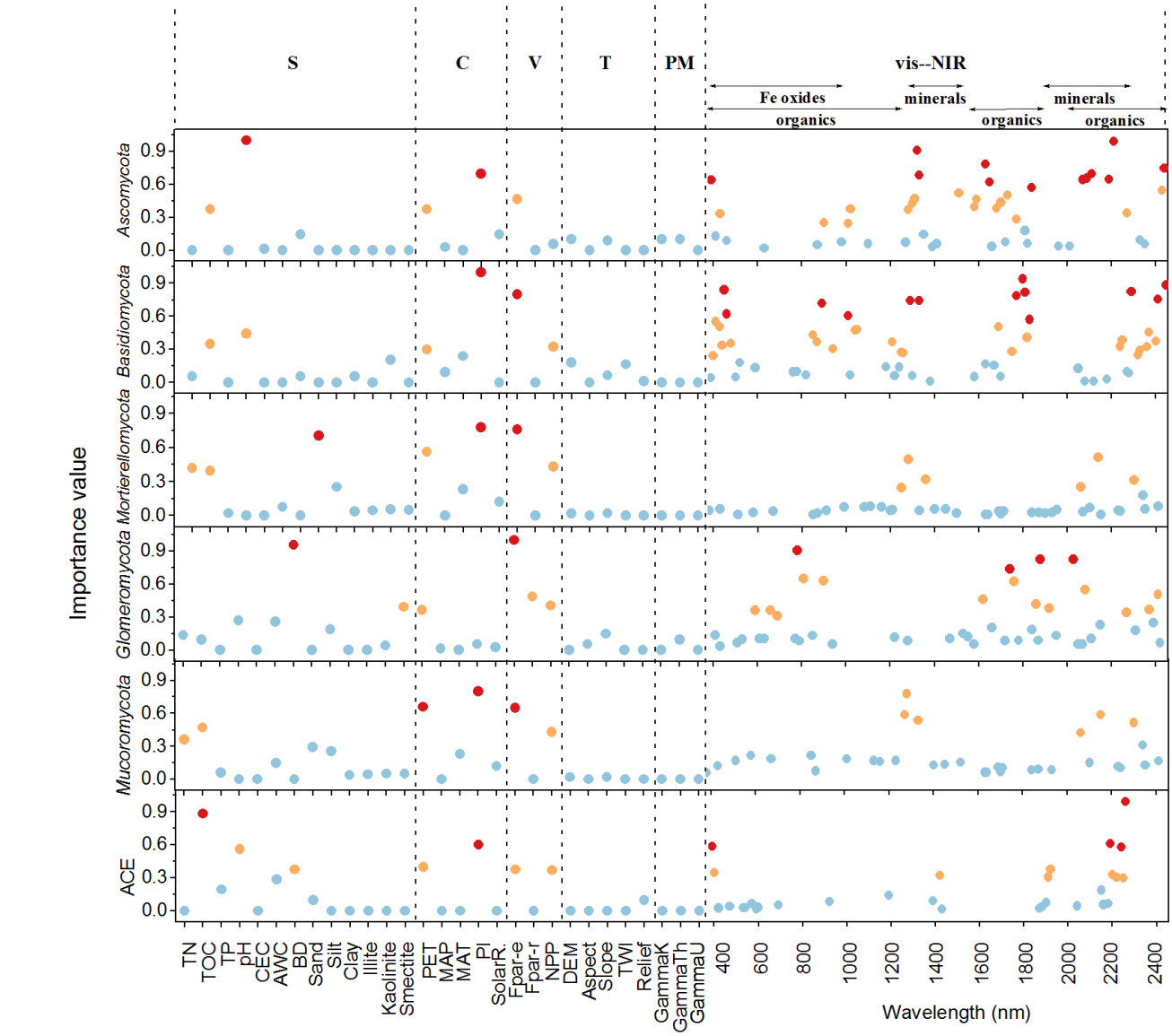

Figure 5. Important predictors of relative abundance of fungal phyla and diversity index measured by the variable importance of the 1D-CNNs spectro-transfer functions (n = 577) derived with publicly accessible data that represent soil (S), climate (C), vegetation (V), terrain (T), parent material (PM) and visible–near infrared (vis–NIR) spectra. The dots in red, orange, and blue color indicated the most, medium, and least important level. The importance value for the majority of wavelengths were low and close to zero value, thus these wavelengths were not shown to make the figure clearer.

The identified wavelengths mostly coincided with absorptions that are related to carbon functional groups found in organic matter, including C-H, N-H, C-O, with a smaller number of wavelengths coinciding with those that are related to clay minerals

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

and Fe-oxides (Table 2). The organic functional groups, C-H alkyl and methyls, N-H of amines, and C-O of carbohydrates, which might indicate the presence of relatively labile forms of carbon, and were important in the models of fungal phyla but not of ACE diversity. The C=O of amides and carboxylic acids, which represent stable forms of carbon were not as important in modeling (Fig. 5). Other wavelengths that represent Fe-oxides and clay minerals were also important in the models, indicating the different ecological niches and physiological characteristics (Table 2).

Table 2. Absorption band assignment for the most important vis–NIR wavelengths in the 1D-CNN models. The assignment of vis–NIR absorptions from Viscarra Rossel and Behrens (2010); Stenberg et al. (2010).

	ACE	Ascomycota	Basidiomycota	Mortierellomycota	Glomeromycota	Mucoromycota
Fe-oxides	390	390	410, 460			
Clay minerals	2190, 2240	1330, 2190, 2210	1330, 2140	1360, 2140		1330, 2150
Organics						
C-H of aromatics		1630, 1650				
N-H of amine		2070, 2090, 2110	1010	2060	780, 2030	2060
C-H of alkyl asymmetric-symmetric doublet			890, 1290	1250, 1280	1740	1270, 1280
C=O of carboxylic acids						
C=O of amides						
C-H of aliphatics						
C-H of methyls		1840, 2440	1770, 1800, 1810 1830, 2450		1880	
C-OH of phenolics						
C-O of carbohydrates	2260		2410, 2290	2300		2300

Other soil properties, such as total organic carbon and pH were important variables in the spectro-transfer functions of Ascomycota and Basidiomycota, and fungal diversity. Total organic carbon and total nitrogen were important in the spectro-transfer functions of Mortierellomycota and Mucoromycota and bulk density was important in the spectro-transfer functions of Glomeromycota, Ascomycota and ACE diversity (Fig. 5). As well as soil properties, climatic factors such as the PI and PET, and vegetation, represented by Fpar-e and NPP were also important in the modelling of fungal phyla relative abundance and community diversity. The variables that we used to represent terrain, and parent material exerted less influence in the models (Fig. 5).

4 Discussion

Soil fungi play essential and diverse functional roles in ecosystem. However, they are challenging to investigate due to laborious, time-consuming and costly field sampling, and laboratory analysis. A paucity in the availability of soil microbial data is thought to be one of the main contributors to the uncertainty of soil health assessment and ecosystem management. Here, we show that spectro-transfer functions with readily accessible vis–NIR spectra and publicly available soil and environmental data can be developed to estimate soil fungal abundance and diversity. Our approach provides a new opportunity to infer the

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

continuous distribution of soil fungal community at a large area with the less sampling density and consequently less laboratory analysis costs. The approach will complement molecular approaches for the assessment, characterization and improved understanding of soil fungal communities and their associated functions at different scales (Hart et al., 2020).

Out of the seven statistical and machine learning models tested, the optimised 1D-CNNs were the most successful for estimating fungal phyla relative abundance and diversity, consistently producing the highest cross-validation R^2 values. The reason might be that the 1D-CNNs can automatically 'learn' the non-linear and complex relations between the soil fungal variables and the covariables. The models extract large features during convolution and adjust the weights of each covariate during the model iterations, which are also back-propagated (Breiman, 2001b; Lecun et al., 2015). Although 1D-CNNs have been used for the spectroscopy modeling of soil physicochemical properties (Ng et al., 2019; Tsakiridis et al., 2020; Shen and Viscarra Rossel, 2021), to our best knowledge, this present study is the first to develop spectro-transfer functions for estimating soil fungal abundance and diversity.

Our results shown that the 1D-CNN spectroscopic models (with only vis–NIR spectra) could explain, on average, 40% of the variation in the relative abundance of fungal phyla and community diversity (R^2 values of 0.30–0.45). It is because these spectra characterise the soil's organic and mineral composition, which serves to supply energy and the elements that fungi use to promote vital activities (Müller, 2015). Microbial activities are closely associated with the types and amounts of organic matter and our results indicate that the most important vis–NIR wavelengths in the modelling of fungal relative abundance and community diversity corresponds to functional groups in the different types of organic compounds in the soils (Viscarra Rossel and Hicks, 2015) (Fig. 5 and Table S2 in Supplementary information).

The 1D-CNN spectro-transfer functions (with vis–NIR spectra and other soil and environmental data) improved the modelling. This suggests that other variables that represent climate, soil nutrients, pH, vegetation, are important predictors of fungal growth. Their use in the spectro-transfer functions provided additional and supplementary information for the modelling. On average, these models could explain 60% of the variation in abundance of fungal phyla relative abundance and diversity (R^2 values of 0.45–0.73).

The soil organic and mineral composition, represented by the vis–NIR spectra, were the most important predictors in the models for fungal relative abundance and community diversity. Additionally, total organic carbon and pH were important predictors of fungal diversity and the relative abundance of Ascomycota and Basidiomycota. Although most soil fungi do not require strict pH ranges for habitation and growth (Rousk et al., 2009; Zhao and Shen, 2018), some basophilic or acidophilic fungi are sensitive to changes in pH (Gai et al., 2006) and saprophytic fungi are thought to be more sensitive to soil pH, compared to other fungi (Kivlin and Hawkes, 2016). Soil bulk density was an important predictor of fungal diversity and the relative abundance of Glomeromycota. Many fungi, including those that form arbuscular mycorrhiza, such as Glomeromycota, infect plants roots achieving mutualistic symbiosis (Schubler et al., 2001). Denser soil bulk density could reduce the availability of soil nutrients and water, leading to poor development of plant roots and a smaller infection rate for the symbiosis. The PI and evapotranspiration were the most important climatic predictors of fungal abundance and diversity in the models. PI represents the soil-water balance which has been shown to affect soil microbial growth at various studies (Bachar et al., 2010; Blankinship et al., 2011; Maestre et al., 2015; Delgadobaquerizo et al., 2018b). Because soil-water stress could strongly restrict microbial

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

activity and distribution by controlling the availability of soil nutrients, pH and oxygen (Delgadobaquerizo et al., 2018b). NPP and Fpar-e were important predictors of fungal diversity and the relative abundance of the five dominant phyla. Larger values of NPP and Fpar-e occur due to greater biomass production and thus more accumulation of litter and coarse organic matter in soil. Soil fungi are some of the decomposers of litter and soil organic matter, including cellulose and lignin, which are often resistant to bacterial decomposition (Treseder and Lennon, 2015; Nicolas et al., 2019).

5 Conclusions

Deep learning with optimised 1D-CNNs provides a new approach to estimate the relative abundance and diversity of soil fungal communities. The 1D-CNNs outperformed the six other machine learning methods tested for estimating the relative abundance of fungal phyla and diversity. The 1D-CNN spectro-transfer (vis–NIR spectra and other soil and environmental data) functions produced more accurate estimates (R^2 0.45–0.73) compared to the spectroscopic (vis–NIR spectra only) models (R^2 0.36–0.55) and models with the soil and environmental data only (R^2 0.38–0.60). As well as the soil organic and mineral composition, represented by vis–NIR spectra, other edaphic, climatic, and biotic factors including soil nutrients, pH, bulk density, potential evapotranspiration, the soil-water balance and net primary productivity were important predictors in the modelling. Given the crucial role of fungi in the functioning of ecosystem, our study helps the development of methods to supplement molecular approaches for a better understanding of the diversity and biogeography of soil fungi over large scales.

Code availability. The code used for the analyses presented in this work is available from the corresponding author on reasonable request.

Data availability. The BASE data are available online at http://www.bioplatforms.com/soil-biodiversity/ after registration.

Author contributions. R.A.V.R conceived the research and designed the study with Y.Y and Z.S. Y.Y and Z.S. carried out the experiments and with R.A.V.R. drafted the manuscript. A.B. derived the data on fungal relative abundance and diversity and edited the manuscript. All authors discussed and interpreted the results and produced the final manuscript.

Competing interests. The authors have no competing interests to declare.

Acknowledgements. This work was performed with funding from Curtin University. We acknowledge the contribution of the Biomes of Australian Soil Environments (BASE) consortium in the generation of data used in this publication. DOI 10.1186/s13742-016-0126-5. The BASE project is supported by funding from Bioplatforms Australia through the Australian Government National Collaborative Research

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

Infrastructure Strategy (NCRIS). The development of the convolutional neural network models was supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

References

Bachar, A., Al-Ashhab, A., Soares, M. I., Sklarz, M. Y., Angel, R., Ungar, E. D., and Gillor, O.: Soil microbial abundance and diversity along a low precipitation gradient, Microbial Ecology, 60, 453–461, 2010.

Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., De Wit, P., Sánchez-García, M., Ebersberger, I., de Sousa, F., Amend, A., Jumpponen, A., Unterseher, M., Kristiansson, E., Abarenkov, K., Bertrand, Y. J. K., Sanli, K., Eriksson, K. M., Vik, U., Veldre, V., and Nilsson, R. H.: Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data, Methods in Ecology and Evolution, 4, 914–919, https://doi.org/https://doi.org/10.1111/2041-210X.12073, 2013.

Bissett, A., Fitzgerald, A., Meintjes, T., Mele, P. M., Reith, F., Dennis, P. G., Breed, M. F., Brown, B., Brown, M. V., Brugger, J., et al.: Introducing BASE: the Biomes of Australian Soil Environments soil microbial diversity database, GigaScience, 5, 21, 2016.

Blankinship, J. C., Niklaus, P. A., and Hungate, B. A.: A meta-analysis of responses of soil biota to global change, Oecologia, 165, 553–565, 2011.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001a.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001b.

Clark, R. N. and Roush, T. L.: Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications, Journal of Geophysical Research, 89, 6329–6340, 1984.

Davinic, M., Fultz, L. M., Acosta-Martinez, V., Calderón, F. J., Cox, S. B., Dowd, S. E., Allen, V. G., Zak, J. C., and Moore-Kucera, J.: Pyrosequencing and mid-infrared spectroscopy reveal distinct aggregate stratification of soil bacterial communities and organic matter composition, Soil Biology & Biochemistry, 46, 63–72, 2012.

Delgadobaquerizo, M., Maestre, F. T., Reich, P. B., Jeffries, T. C., Gaitan, J. J., Encinar, D., Berdugo, M., Campbell, C. D., and Singh, B. K.: Microbial diversity drives multifunctionality in terrestrial ecosystems., Nature Communications, 7, 10 541–10 541, 2016.

Delgadobaquerizo, M., Oliverio, A. M., Brewer, T. E., Benaventgonzalez, A., Eldridge, D. J., Bardgett, R. D., Maestre, F. T., Singh, B. K., and Fierer, N.: A global atlas of the dominant bacteria found in soil, Science, 359, 320–325, 2018a.

Delgadobaquerizo, M., Reith, F., Dennis, P. G., Hamonts, K., Powell, J. R., Young, A. G., Singh, B. K., and Bissett, A.: Ecological drivers of soil microbial diversity and soil biological networks in the Southern Hemisphere, Ecology, 99, 583–596, 2018b.

Donohue, R. J., McVicar, T., and Roderick, M. L.: Climate-related trends in Australian vegetation cover as inferred from satellite observations, 1981–2006, Global Change Biology, 15, 1025–1039, 2009.

Fisher, A., Rudin, C., and Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously., Journal of Machine Learning Research, 20, 1–81, 2019.

Friedman, J. H.: Greedy function approximation: A gradient boosting machine., Annals of Statistics, 29, 1189–1232, 2001.

Gai, J., Christie, P., Feng, G., and Li, X. L.: Twenty years of research on community composition and species distribution of arbuscular mycorrhizal fungi in China : a review, Mycorrhiza, 16, 229–239, 2006.

Gallant, J., Wilson, N., Dowling, T., Read, A., and Inskeep, C.: SRTM-derived 1 second digital elevation models version 1.0, Geoscience Australia: Canberra, ACT, 2011.

Griffiths, R. I., Thomson, B. C., James, P., Bell, T., Bailey, M., and Whiteley, A. S.: The bacterial biogeography of British soils, Environmental Microbiology, 13, 1642–1654, 2011.

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

Hart, M. M., Cross, A. T., D'Agui, H. M., Dixon, K. W., Van der Heyde, M., Mickan, B., Horst, C., Grez, B. M., Valliere, J. M., Viscarra Rossel, R. A., Whiteley, A., Wong, W. S., Zhong, H., and Nevill, P.: Examining assumptions of soil microbial ecology in the monitoring of ecological restoration, Ecological Solutions and Evidence, 1, e12 031, https://doi.org/https://doi.org/10.1002/2688-8319.12031, 2020.

Haverd, V., Raupach, M. R., Briggs, P. R., Canadell, J. G., Isaac, P., Pickett-Heaps, C., Roxburgh, S. H., van Gorsel, E., Viscarra Rossel, R. A., and Wang, Z.: Multiple observation types reduce uncertainty in Australia's terrestrial carbon and water cycles, Biogeosciences, 10, 2011–2040, https://doi.org/10.5194/bg-10-2011-2013, 2013.

Jenny, H.: Factors of Soil Formation: A System of Quantitative Pedology, Courier Corporation, New York, 1994.

Kivlin, S. N. and Hawkes, C. V.: Tree species, spatial heterogeneity, and seasonality drive soil fungal abundance, richness, and composition in Neotropical rainforests, Environmental Microbiology, 18, 4662–4673, 2016.

Kuhn, M., Leeuw, J. D., and Zeileis, A.: Building Predictive Models in R Using the caret Package, Journal of Statistical Software, 28, 1–26, 2008.

Lecun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, https://doi.org/10.1038/nature14539, 2015.

Li, J., Delgadobaquerizo, M., Wang, J., Hu, H., Cai, Z., Zhu, Y., and Singh, B. K.: Fungal richness contributes to multifunctionality in boreal forest soil, Soil Biology & Biochemistry, 136, 107 526, 2019.

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

Lozupone, C. A. and Knight, R.: Species divergence and the measurement of microbial diversity, FEMS Microbiology Reviews, 32, 557–578, 2008.

Maestre, F. T., Delgadobaquerizo, M., Jeffries, T. C., Eldridge, D. J., Ochoa, V., Gozalo, B., Quero, J. L., Garciagomez, M., Gallardo, A., Ulrich, W., et al.: Increasing aridity reduces soil microbial diversity and abundance in global drylands., Proceedings of the National Academy of Sciences of the United States of America, 112, 15 684–15 689, 2015.

Minty, B., Franklin, R., Milligan, P., Richardson, M., and Wilford, J.: The radiometric map of Australia, Exploration Geophysics, 40, 325–333, 2009.

Morellos, A., Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., and Mouazen, A. M.: Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy, Biosystems Engineering, 152, 104–116, 2016.

Müller, B.: Experimental interactions between clay minerals and bacteria: a review, Pedosphere, 25, 799–810, 2015.

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

Nicolas, C., Martinbertelsen, T., Floudas, D., Bentzer, J., Smits, M. M., Johansson, T., Troein, C., Persson, P., and Tunlid, A.: The soil organic matter decomposition mechanisms in ectomycorrhizal fungi are tuned for liberating soil organic nitrogen, The ISME Journal, 13, 977–988, 2019.

Nilsson, R. H., Larsson, K.-H., Taylor, A. F., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F. O., Tedersoo, L., Saar, I., Kõljalg, U., and Abarenkov, K.: The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications, Nucleic Acids Research, 47, D259–D264, https://doi.org/10.1093/nar/gky1022, 2018.

Prescott, J. A.: A climatic index for the leaching factor in soil formation, Journal of Soil Science, 1, 9–19, 1950.

Quinlan, J. R.: Learning with continuous classes, in: 5th Australian joint conference on artificial intelligence, vol. 92, pp. 343–348, Singapore, 1992.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org, 2014.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2005.

Rousk, J., Brookes, P. C., and Baath, E.: Contrasting Soil pH Effects on Fungal and Bacterial Growth Suggest Functional Redundancy in Carbon Mineralization, Applied and Environmental Microbiology, 75, 1589–1596, 2009.

Savitzky, A. and Golay, M. J. E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures., Analytical Chemistry, 36, 1627–1639, 1964.

Schubler, A., Schwarzott, D., and Walker, C.: A new fungal phylum, the Glomeromycota: phylogeny and evolution, Fungal Biology, 105, 1413–1421, 2001.

Serna-Chavez, H. M., Fierer, N., and Bodegom, P. M.: Global drivers and patterns of microbial abundance in soil, Global Ecology and Biogeography, 22, 1162–1172, 2013.

Shen, Z. and Viscarra Rossel, R. A.: Automated spectroscopic modelling with optimised convolutional neural networks, Scientific Reports, 11, 208, https://doi.org/10.1038/s41598-020-80486-9, 2021.

Shi, Z., Ji, W., Viscarra Rossel, R. A., Chen, S., and Zhou, Y.: Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis–NIR spectral library, European Journal of Soil Science, 66, 679–687, 2015.

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., and Wetterlind, J.: Visible and Near Infrared Spectroscopy in Soil Science, Advances in Agronomy, 107, 163–215, 2010.

Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D., and Vandewalle, J.: Least Squares Support Vector Machines, 2002.

Treseder, K. K. and Lennon, J. T.: Fungal Traits That Drive Ecosystem Dynamics on Land, Microbiology and Molecular Biology Reviews, 79, 243–262, 2015.

Tsakiridis, N. L., Keramaris, K. D., Theocharis, J. B., and Zalidis, G. C.: Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network, Geoderma, 367, 114 208, https://doi.org/10.1016/j.geoderma.2020.114208, 2020.

Vetrovsky, T., Kohout, P., Kopecky, M., Machac, A., Man, M., Bahnmann, B. D., Brabcova, V., Choi, J., Meszarosova, L., Human, Z. R., et al.: A meta-analysis of global fungal distribution reveals climate-driven patterns., Nature Communications, 10, 5142, 2019.

Viscarra Rossel, R., Bui, E., Caritat, P., and N.J., M.: Mapping iron oxides and the color of Australian soil usingvisible–near-infrared reflectance spectra, Journal OF Geophysical Research, 115, F04 031, 2010.

Viscarra Rossel, R. A.: Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra, Journal of Geophysical Research: Earth Surface, 116, F04 023, 2011.

Viscarra Rossel, R. A. and Behrens, T.: Using data mining to model and interpret soil diffuse reflectance spectra, Geoderma, 158, 46–54, 2010.

Viscarra Rossel, R. A. and Brus, D. J.: The cost-efficiency and reliability of two methods for soil organic C accounting, Land Degradation & Development, 29, 506–520, 2018.

Viscarra Rossel, R. A. and Hicks, W. S.: Soil organic carbon and its fractions estimated by visible-near infrared transfer functions, European Journal of Soil Science, 66, 438–450, https://doi.org/10.1111/ejss.12237, 2015.

https://doi.org/10.5194/soil-2021-79
Preprint. Discussion started: 13 September 2021

Viscarra Rossel, R. A. and McBratney, A. B.: Soil chemical analytical accuracy and costs: implications from precision agriculture., Australian Journal of Experimental Agriculture, 38, 765–775, 1998.

Viscarra Rossel, R. A., Chen, C., Grundy, M., Searle, R., Clifford, D., and Campbell, P.: The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project, Soil Research, 53, 845–864, 2015.

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D., 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., and Ji, W.: A global spectral library to characterize the world's soil, Earth-Science Reviews, 155, 198–230, https://doi.org/10.1016/j.earscirev.2016.01.012, 2016.

Wold, S., Sjostrom, M., and Eriksson, L.: PLS-REGRESSION: A BASIC TOOL OF CHEMOMETRICS, Chemometrics and Intelligent Laboratory Systems, 58, 109–130, 2001.

Xu, T. and Hutchinson, M.: ANUCLIM version 6.1 user guide, The Australian National University, Fenner School of Environment and Society, Canberra, 2011.

Yang, Y., Viscarra Rossel, R., Li, S., Bissett, A., Lee, J., Shi, Z., Behrens, T., and Court, L. N.: Soil bacterial abundance and diversity better explained and predicted with spectro-transfer functions, Soil Biology & Biochemistry, 129, 29–38, 2019.

Zhao, X. Q. and Shen, R. F.: Aluminum-Nitrogen Interactions in the Soil-Plant System., Frontiers in Plant Science, 9, 807, 2018.